JANUARY 2021

AI and the Future of Cyber Competition

CSET Issue Brief

AUTHOR

Wyatt Hoffman

Table of Contents

Executive Summary

As artificial intelligence begins to transform cybersecurity, the pressure to adapt may put competing states on a collision course. Recent advances in machine learning techniques could enable groundbreaking capabilities in the future, including defenses that automatically interdict attackers and reshape networks to mitigate offensive operations. Yet even the most robust machine learning cyber defenses could have potentially fatal flaws that attackers can exploit. Rather than end the cat-and-mouse game between cyber attackers and defenders, machine learning may usher in a dangerous new chapter.

Could embracing machine learning systems for cyber defense actually exacerbate the challenges and risks of cyber competition? This study aims to demonstrate the possibility that machine learning could shape cyber operations in ways that drive more aggressive and destabilizing engagements between states. While this forecast is necessarily speculative, its purpose is practical: to anticipate how adversaries might adapt their tactics and strategies, and to determine what challenges might emerge for defenders. It derives from existing research demonstrating the challenges machine learning faces in dynamic environments with adaptive adversaries.

This study envisions a possible future in which cyber engagements among top-tier actors come to revolve around efforts to target attack vectors unique to machine learning systems or, conversely, defend against attempts to do so. These attack vectors stem from flaws in machine learning systems that can render them susceptible to deception and manipulation. These flaws emerge because of how machine learning systems "think," and unlike traditional software vulnerabilities, they cannot simply be patched. This dynamic leads to two propositions for how these attack vectors could shape cyber operations.

The first proposition concerns offense: Attackers may need to intrude deep into target networks well in advance of an attack in order to circumvent or defeat machine learning defenses. Crafting an attack that can reliably deceive a machine learning system requires knowing a specific flaw in how the system thinks. But discovering such a flaw may be difficult if the system is not widely exposed or publicly available. To reach a hardened target, an attacker may try to compromise the system during development. An attacker with sufficient access could reverse-engineer a system during its development to discover a flaw or even create one by sabotaging the process. This opportunity to gain intelligence on an adversary's defenses creates more value in intruding into adversary computer networks well in advance of any planned attack.

The second proposition concerns defense: Guarding against deceptive attacks may demand constant efforts to gain advanced knowledge of attackers' capabilities. Because machine learning systems cannot simply be patched, they must be able to adapt to defend against deceptive attacks. Yet researchers have found that adaptations to defend against one form of deception are vulnerable to another form of deception. No defense has been found that can make a machine learning system robust to all possible attacks—and it is possible none will be found. Consequently, machine learning systems that adapt to better defend against one form of attack may be at constant risk of becoming vulnerable to another. In the face of an imminent threat by an adversary, the best defense may be to intrude into the adversary's networks and gain information to harden the defense against their specific capabilities.

Together these two propositions suggest machine learning could amplify the most destabilizing dynamics already present in cyber competition. Whether attacking or defending, at the top tier of operations, machine learning attack vectors may create challenges best resolved by intruding into a competitor's networks to acquire information in advance of an engagement. This would add to existing pressures on states to hack into their adversaries' networks to create offensive options and protect critical systems against adversaries' own capabilities. Yet the target of an intrusion may view the intrusion as an even greater threat—regardless of motive—if it could reveal information that compromised machine learning defenses. The already blurred line between offensive and defensive cyber operations may fade further. In a crisis, the potential for cyber operations to accelerate the path to conflict may rise. In peacetime, machine learning may fuel the steady escalation of cyber competition. Adversaries may adapt by targeting machine learning itself, including:

- Compromising supply chains or training processes to insert backdoors into machine learning systems that expose a potentially wide swath of applications to possible attacks.
- Poisoning training data, such as open source malware repositories, to degrade cybersecurity applications.
- Unleashing risky capabilities to circumvent defenses, such as malware with greater degrees of autonomy.
- Targeting defenders' trust in machine learning systems, such as by inducing systems to generate "false positives" by mislabeling legitimate files as malware.

For the United States and its allies, harnessing machine learning for cybersecurity depends on anticipating and preparing for these potential changes to the threat landscape. If cyber defense increasingly relies on inherently flawed machine learning systems, frameworks and metrics will be needed to inform risk-based decisions about where and how to employ them. Securing the machine learning supply chain will demand collective governmental and private sector efforts. Finally, the United States and its allies must exercise caution in the conduct of their offensive operations and communicate with adversaries to clarify intentions and avoid escalation.

Introduction

States seeking competitive advantage will likely turn to artificial intelligence to gain an edge in cyber conflict. Cybersecurity ranks high among priority applications for those leading AI development.[1] China and Russia see in AI the potential for decisive strategic advantage.[2] Military planners in the United States envision systems capable of automatically conducting offensive and defensive cyber operations.[3] On the precipice of a potential collision between AI competition and cyber conflict, there is still little sense of the potential implications for security and stability.

AI promises to augment and automate cybersecurity functions. Network defenders have already begun to reap the benefits of proven machine learning methods for the data-driven problems they routinely face.[4] Even more tantalizing is the speculative prospect of harnessing for cybersecurity the cutting-edge machine learning techniques that yielded "superhuman" performance at chess and the Chinese board game Go.

Yet the machine learning capabilities fueling these applications are no panacea. These systems often suffer from inherent flaws. Unlike traditional software vulnerabilities, these flaws emerge because of how these systems make inferences from data—or, more simply, how they "think." These flaws can lead even highly robust systems to fail catastrophically in the face of unforeseen circumstances. In the race between researchers developing ways to safeguard these systems and those seeking to break them, the attackers appear to be winning.

What will happen when these powerful yet flawed machine learning capabilities enter into the dynamic, adversarial context of cyber competition? Machine learning can help mitigate traditional cyber attack vectors, but it also creates new ones that target machine learning itself. Attackers will systematically try to break these systems. A growing body of technical research explores machine learning attack vectors and prospective defenses. Yet there has been little effort to analyze how these changes at a technical level might impact cyber operations and, in turn, their strategic dynamics.

This study approaches this problem by exploring a possible worst-case scenario: machine learning could amplify the most destabilizing dynamics already present in cyber competition. The purpose is not to lay out a case against harnessing machine learning for cybersecurity. Precisely because these capabilities could become crucial to cyber defense, the aim here is to provoke thinking on how to proactively manage the geopolitical implications of persistent technical flaws.

This study explores how the attack vectors unique to machine learning might change how states hack each other's critical networks and defend their own. The unique vulnerabilities of these systems may create problems for both offense and defense best resolved by intruding into adversaries' systems in advance of an engagement. For offense, this arises from the potential need for exquisite intelligence on, or even direct access to, a machine learning system to reliably defeat it. For defense, this arises from the need for advanced knowledge of a specific attack methodology to ensure a defense's viability against it.

The combination of these offensive and defensive imperatives could exacerbate the escalation risks of cyber engagements. States would have even stronger incentives to intrude into one another's systems to maintain offensive options (for contingencies such as armed conflict or strategic deterrence) and to ensure the viability of their own defenses. Yet it may be even harder to differentiate cyber espionage from intrusions laying the groundwork for an attack; the target of an intrusion may assume that it is preparation for an imminent attack, or that it will at the very least enable offensive options. As adversaries struggle to gain an edge over one another, the line between offense and defense—tenuous as it already is in cyber operations—fades. This dynamic may fuel the steady drumbeat of cyber competition in peacetime. In a crisis, the potential for misinterpretation of a cyber operation to trigger escalation may rise.

This forecast rests on two core assumptions that must be addressed at the outset. These are certainly debatable but the aim is to analyze their implications should they hold, not assess how likely they are to do so.

The first is that machine learning could plausibly deliver on the promise of sophisticated, automated cyber defenses at scale. That is, the significant technical and practical hurdles (e.g., demands for high quality data and computing power, as well as organizational challenges to implementation) will not prove insurmountable at least for top-tier actors such as China and the United States. This study begins with a survey of applications in various stages of development to demonstrate their plausibility. But it makes no attempt to assess the current state of play with deployed machine learning cybersecurity applications or the likelihood of realizing them in the near term.[*]

[*] For a more thorough survey of existing applications and near-term prospects for machine learning in cybersecurity see Micah Musser and Ashton Garriott, "Machine Learning and Cybersecurity: Hype and Reality" (Center for Security and Emerging Technology, forthcoming).

The second assumption is that insights from existing research on machine learning attack vectors will hold at least for the prevailing machine learning methods and applications discussed here. This study draws extensively on research demonstrating the attack vectors targeting machine learning and what these vectors reveal about the potential limits of the robustness of machine learning systems. It makes no assumptions about yet unseen innovations in machine learning techniques or offensive or defensive measures that might fundamentally change the trajectory.

This study begins with a brief overview of machine learning applications for cybersecurity, including their prospective defensive benefits and inherent flaws. It then examines two propositions for how these technical changes to the cybersecurity landscape may, in turn, shape offensive and defensive cyber operations. Specifically, machine learning attack vectors could create predicaments that incentivize intrusions into adversaries' networks, whether to create offensive options or shore up defenses. This study continues on to explore how the combination of these two propositions could fuel the steady intensification of cyber competition and increase the risks of misperception and escalation in cyber engagements.

Promise and Pitfalls of Artificial Intelligence for Cybersecurity

Machine learning lies at the core of the emerging and maturing cybersecurity applications discussed throughout this paper. Described as an approach to, or subfield of, AI, machine learning has fueled recent milestones in tasks ranging from image recognition to speech generation to autonomous driving.

Machine learning systems essentially adapt themselves to solve a given problem.[5] This process often starts with a blank slate in the form of a neural network. The system's developers feed a dataset to the neural network and an algorithm shapes the network's structure to adapt to the patterns within this data. For example, a system for analyzing malware will learn to accurately identify a file as "malware" or "benign" and associate each classification with particular patterns. Eventually the network develops a generalized model of what malware "looks like."

High quality training data, effective training algorithms, and substantial computing power comprise the critical inputs to this process. The resulting machine learning model, ideally, detects not only known malware but yet unseen variants. Advancements in machine learning techniques reduce the need for human experts to structure data.* Rather than relying on an expert to tell the model what key features of malware to look for, the model discovers on its own how to classify malware. As a result, it may find ways of identifying malware more effective at coping with attackers' attempts at obfuscation, such as "metamorphic" malware that rewrites parts of its code as it propagates.[6]

Intrusion detection—finding an adversary's illicit presence in a friendly computer network—may benefit similarly from machine learning. Existing intrusion detection systems already look for red flags, such as a computer becoming active in the middle of the night or a user attempting to access files unrelated to their work. Yet defenders struggle to sort through the vast data generated by network activity in large enterprises, allowing attackers to hide in the noise. Machine learning systems can turn this data into a major

---

* Deep learning architectures are particularly promising in this respect. For example, one approach translates a piece of malware into an image by converting code to pixels in order to utilize advances in deep learning-based image classification as a means of classifying the underlying code as benign or malicious. See Daniel Gibert, Carles Mateu, and Jordi Planes, "The Rise of Machine Learning for Detection and Classification of Malware: Research Developments, Trends and Challenges," *Journal of Network and Computer Applications* 153 (March 1, 2020): 102526.

advantage. By fusing information from a wider and more diverse range of sensors throughout the environment, they create a baseline of normal network activity against which even slight deviations can be detected.[7]

AI and machine learning have quickly become buzzwords in the cybersecurity industry. This makes it difficult to assess the extent to which these capabilities are actually relied upon or are invoked for marketing purposes. Cybersecurity vendors commonly claim to leverage machine learning.* For example, as CrowdStrike defends its customers' devices and networks, it rakes in data on around 250 billion events daily and feeds the data to machine learning models to predict new kinds of attacks.[8] Darktrace states that it employs multiple machine learing methods in its "Enterprise Immune System," empowering systems that can automatically mitigate attacks.[9] Machine learning has also been harnessed to test software for vulnerabilities, detect spam and spear-phishing attacks, and identify suspicious behavior and insider threats.[10] In general, machine learning systems appear to be deployed mainly for relatively narrow tasks in support of human network defenders.[11]

Traditional machine learning methods relying on large training datasets may not suffice for a system that performs more complex tasks requiring sequences of actions, each dependent upon the outcome of the last. Such a system needs to learn more like a human—through experimentation and trial-and-error. This is the essence of *reinforcement learning*. Instead of being fed training data, a reinforcement learning agent interacts with a simulated environment and is rewarded for action that advances its objective. It gradually learns sets of moves, or "policies," to guide its action. The process can yield stunning results, such as the victory by AlphaGo, a reinforcement learning system developed by DeepMind, over Lee Sedol, the world champion in the incredibly complex game of Go.[12]

If reinforcement learning can master chess and Go, it might unlock future cyber defenses capable of discovering and automatically executing moves and strategies in the "game" against cyber attackers. Cyber defenders have a home field advantage.[13] They can change the configuration of networks to interfere with an attacker or deploy decoy systems such as "honeypots" that lure attackers in and lead them to reveal capabilities. However, setting up

---

* According to one survey of U.S., UK, and German businesses 82 percent of respondents stated their company employed a cybersecurity product utilizing machine learning in some form. See Ondrej Kubovič, "Machine-Learning Era in Cybersecurity: A Step Toward a Safer World or the Brink of Chaos?" ESET, February 2019, https://www.eset.com/fileadmin/ESET/US/download/ESETus-Machine-Learning-Era-in-Cybersecurity-Whitepaper-WEB.pdf.

honeypots and reconfiguring networks are technically demanding tasks and, to be effective, require the ability to anticipate an attacker's moves and adapt on the fly.[14] While still largely confined to academic research, and thus more speculative, pioneering applications of reinforcement learning may produce systems capable of these feats.[15] Reinforcement learning agents could learn optimal strategies for reconfiguring networks and mitigating attacks, rapidly analyze an attacker's moves and select and execute actions, such as isolating or patching infected nodes and deploying honeypots. At a minimum, they could present attackers with a constantly moving target, introducing uncertainty and increasing the complexity required for offensive operations.[16]

Machine learning could plausibly deliver on the promise of cyber defenses that adapt to novel threats and automatically engage attackers. These potentially game-changing applications are the focus of this study, even though the most significant near-term gains for cybersecurity may be found in automating the more "mundane" aspects of cybersecurity. The more speculative capabilities may not be realized in the near term, but given their potential to transform cyber operations it is worth exploring their implications.

Security vulnerabilities of machine learning

As promising as they are, most machine learning cyber capabilities have yet to face the most important test: systematic attempts by attackers to break them once deployed. Machine learning can fail catastrophically under certain conditions.[17] Evidence for this includes "adversarial examples": manipulated inputs (often images that have been subtly altered) created by researchers to trick machine learning models. Seemingly imperceptible changes to an image of a turtle can cause a model that otherwise classifies it with perfect accuracy to mistake it for a rifle.[18] Similar adversarial techniques can cause reinforcement learning systems to malfunction.[19]

Adversarial examples reveal a problem inherent to machine learning, not just deficiencies in specific systems. Every model rests on assumptions about the data to make decisions—assumptions, for instance, about what malware "looks like." If an input violates those assumptions it will fool the model (and often a successful deception fools other models trained for the same task).[20] Flawed training methods or data can create vulnerabilities. But models can also become vulnerable when the conditions in which they are deployed change in ways that violate assumptions learned in training. The model's predictions will no longer be accurate—a problem referred to as "concept drift."[21] Even slight deviations from training conditions can dramatically degrade performance.

This poses a constant problem for machine learning applications in dynamic, adversarial contexts like cybersecurity.[22] For machine learning cyber defenses to be viable, they may have to learn and evolve not just during training, but in deployment.[23] Systems will have to keep up with a constantly changing cybersecurity landscape. For instance, an intrusion detection system modeling "normal" network activity must constantly revise this model as legitimate and malicious activity changes. The system might generate new training data by observing the behavior of devices connected to the network, using this data to continuously update and refine its model to better predict future behavior.

Innovative machine learning techniques aim to create systems capable of better contending with adaptive adversaries in dynamic environments. These techniques harness competition to drive evolution. For instance, Kelly et al. co-evolve defenses that automatically reconfigure networks to catch the attackers with offensive agents seeking to evade detection.[24] Developers may pit a reinforcement learning agent against an adversarial agent whose objective is to thwart it.[25] These methods attempt to simulate an "arms race" between attackers and defenders to produce models that better anticipate and preempt attacker moves in the real world.[26]

All of this sets the stage for a potential transformation in the cat-and-mouse game between cyber attackers and defenders. The future cybersecurity playing field may feature defenses that evolve automatically through engagements, but such defenses inevitably create new attack vectors that are difficult to safeguard. The next two sections explore how attackers and defenders alike might adapt to these changing technical conditions, setting the foundation to examine the geopolitical implications that follow.

The Imperatives of Offense

If improved machine learning defenses offer significant benefits to defenders, they will introduce significant new hurdles into the planning and execution of offensive cyber operations. Offensive operations often require careful planning and preparation of the target environment. The presence of sophisticated machine learning defenses may force attackers to shift their efforts toward targeting the underlying machine learning models themselves. But hacking machine learning presents its own unique set of problems. The core challenge for attackers will be figuring out how to reliably manipulate or circumvent these systems.

Attacking machine learning

Attackers tend to follow the path of least resistance. If possible, they will try to avoid machine learning defenses entirely, including by targeting "traditional" attack vectors, such as acquiring credentials via spear-phishing. Avoidance, however, may not always be an option. An attacker may attempt to evade the defensive system by exploiting a weakness in the model. Researchers at security firm Skylight Cyber demonstrated how to do so against Cylance's leading machine learning-based antivirus product.[27] Using publicly accessible information, they reverse-engineered the model to discover how it classified files. In the process, they discovered a bias in the model; it strongly associated certain sequences of characters with benign files. A file that otherwise appeared highly suspicious would still be classified as benign if it contained one of the character sequences. The Skylight researchers discovered, in their words, a "universal bypass"—characters that they could attach to almost any piece of malware to disguise it as a benign file.[*] The researchers found that applying their bypass to a sample of 384 malicious files resulted in the machine learning system classifying 84 percent as "benign," often with high confidence.[28]

Attackers will not always be so lucky as to discover a bypass as readily exploitable as in the Cylance case. They could sabotage a model to similar effect. Injecting bad samples into a training dataset (e.g. malware labeled as "benign") can "poison" a model. Even an unsophisticated poisoning attack could dramatically reduce the model's performance.[29] More insidiously, an

---

[*] Cylance disputed the characterization as a "universal bypass" and claimed to have fixed the flaw shortly after being made aware by Skylight. See "Resolution for BlackBerry Cylance Bypass," *BlackBerry ThreatVector Blog*, July 21, 2019, https://blogs.blackberry.com/en/2019/07/resolution-for-blackberry-cylance-bypass.

attacker could poison a model so that it reacts to specific inputs in a way favorable to the attacker—inserting a "backdoor" into the model. In one demonstration, researchers created a "watermark" in the form of a specific set of features in a file that functioned similar to the bypass discovered by Skylight. By tampering with just one percent of the training data, they could induce a model to misclassify malicious files containing the watermark as benign with a 97 percent success rate.[30]

While these examples describe attacks on classification systems, reinforcement learning agents engaged in more complex tasks have similarly proven susceptible to evasion and sabotage.[31] For example, an attacker could poison a defensive system that automatically reconfigures networks so that it responds poorly in specific circumstances; the attacker might trick the system into connecting an infected node to others in the network, rather than isolating it.[32]

The attacker's predicament

The feasibility of evading or poisoning a machine learning system will inevitably depend on the context. It's one thing to demonstrate attacks on machine learning in experimental settings, but it's another to execute them in the real world against a competent defender. In the Cylance case, the attackers benefited from insights into the inner workings of the model. States seeking to create and sustain offensive options may face strategic targets that are not so widely exposed. The difficulty of conducting attacks on machine learning systems under realistic constraints may pressure states to intrude into adversaries' networks to begin laying the groundwork for attacks as early as possible. This pressure stems from the necessity intrusions play in enabling the kinds of attacks described above:

(1) Acquiring information to craft more reliable and effective evasion attacks against machine learning systems: As Goodfellow et al. observe, the greater the attacker's "box knowledge"—knowledge of the target model parameters, architecture, training data and methods—the easier it is to construct an attack that defeats the system.[33] Under "white box" conditions, where the attacker has complete knowledge, crafting an attack is a relatively straightforward matter of optimizing the features of malware (or other inputs) to exploit the model's assumptions.*

* Researchers have naturally found greater success evading antivirus systems and attacking reinforcement learning policies with white-box attacks than with black-box attacks. See, for instance, Hyrum S. Anderson et al., "Learning to Evade Static PE Machine Learning Malware

"Black box" attacks, where the attacker has little to no knowledge of the target model, are possible, but require more guesswork. The attacker may engineer an attack against a substitute for the target model in the hopes that if it fools the substitute, it will fool the target. But this depends on how closely the substitute matches the target.[34] Demonstrations of black-box attacks often leverage publicly available details or the ability to repeatedly probe a target model in order to derive information on how it works. An attacker might buy a commercial service to gain insights into a model, allowing greater flexibility to craft attacks. In top-tier cyber competition, however, an attacker may not enjoy these advantages. If the target model is not widely exposed, attempts to probe it may tip off the defender. And gaining information on some types of defenses, like those that reconfigure networks, would require intruding into the network. Moreover, future security measures may prevent deployed machine learning systems from "leaking" useful information to an attacker attempting to probe them.[35] The best way to acquire box knowledge, then, may be to gain access to a training environment and steal training data or even a trained model.

(2) Compromising systems to enable future exploitation: It is possible to undermine a deployed model, for instance interacting with an intrusion detection system to "normalize" an intruder's presence to it.[36] But competent defenders will be alert to the possibility. The development process may present a softer target.[37] Rather than a model developed from scratch, many applications take existing pre-trained models and tailor them for specific tasks through additional training and fine-tuning in a process known as transfer learning. A backdoor inserted into the pre-trained model can make its way into subsequent models derived from it.[38] This opens up new attack vectors. For example, compromising an open source project, code repository, or a commercial contractor assisting with the development of cybersecurity applications may allow an attacker to insert vulnerabilities deep into systems that make their way into more tightly-controlled training environments. Targeting the development process has the added benefit of scalability: inserting a backdoor into one model may facilitate access to a wide swath of subsequent targets. A transfer learning service supporting diverse commercial, military, or other national security-relevant applications would be a tempting target.

Models via Reinforcement Learning," *arXiv [cs.CR]* (January 26, 2018), arXiv, http://arxiv.org/abs/1801.08917; Huang et al., "Adversarial Attacks on Neural Network Policies."

The Imperatives of Defense

As attackers adapt to the deployment of machine learning, the success or failure of cyber defenses may hinge on the security of machine learning models against deception and manipulation. Yet it has proven difficult to create machine learning systems that are truly robust—that is, systems that can contend with attackers that adapt their tactics to try and defeat them. Innovative defenses against the kinds of attacks described above have emerged, but are routinely broken. Some experts question whether progress toward truly robust machine learning has been illusory.[39] The core challenge for defenders may be safeguarding systems with inherent flaws baked in.

The perpetual problem of machine learning robustness

When a vulnerability is discovered, a machine learning model cannot simply be patched like traditional software. Instead, the developer must retrain the model using adversarial examples or certain training procedures designed to make the model more robust to a particular set of deceptive inputs. However, adjusting the model may simultaneously make it more robust to one set of deceptions but more susceptible to others. Two prominent machine learning security experts, Ian Goodfellow and Nicolas Papernot, thus characterize existing defensive measures as "playing a game of whack-a-mole: they close some vulnerabilities, but leave others open."[40] Such were the findings of Tramer et al., who systematically defeated 13 defenses shown to be effective against adaptive attackers.[41] A similar phenomenon has been observed with reinforcement learning agents; rather than becoming generally robust, those trained against an adaptive adversary in simulated games tend to "overfit" to the adversary. In other words, their adaptations to deal with the regular opponent can leave them vulnerable to a novel attack.[42]

The ease with which defenses are broken may simply reflect the nascent state of machine learning security. But it suggests a more concerning possibility: no defense will be robust to all possible attacks. As David Brumley puts it: "for any ML algorithm, an adversary can likely create [an attack] that violates assumptions and therefore the ML algorithm performs poorly."[43] Unlike software security, which is, at least in theory, a "linear" process of improvement as the developer tests, patches, and repeats, machine learning may present a perpetual security problem. The system can be hardened to any known attack but may always be vulnerable to a possible novel attack.

These observations raise two questions regarding the potential limits on machine learning robustness:

First, how much of a problem do machine learning's flaws pose for the defender? With sufficiently comprehensive training data to accurately model threats, perhaps the risk of a novel attack defeating the system would be negligible. However, cybersecurity presents a uniquely difficult deployment context: Threats continuously evolve, so a deployed system must constantly take in new data to adapt. But if instead of becoming generally robust, machine learning defenses are just playing whack-a-mole, then there may always be an attack that breaks them. Testing systems to try and discover every flaw may prove futile because of the vast range of possible moves the attacker could make to deceive the machine learning model.[44] And attackers may be in a position where they could feasibly discover flaws by repeatedly probing defenses, unlike other domains where engagements between attackers and defenders might be episodic (e.g. autonomous weapon systems in kinetic warfare).*

Second, is this problem endemic to machine learning or a limitation of prevailing methods? It is at least possible that the limits on robustness prove persistent in contexts where systems have to evolve with adaptive adversaries. The process of neural network evolution drives toward efficient solutions to problems, not necessarily solutions that are robust against adaptive adversaries. In the Cylance case, the system discovered an efficient way to classify the whitelisted files—but one that attackers could exploit. This may not matter in some contexts, but systems forced to co-evolve with adaptive adversaries may adapt in ways that inevitably create vulnerabilities. Colbaugh and Glass thus argue that systems that co-evolve with adaptive adversaries become "robust yet fragile."[45] They become effective at dealing with recurrent threats but, in adapting to do so, develop "hidden failure modes" that a novel attack could trigger. Consequently, they argue, prospective mitigations like "ensemble" models, which combine multiple algorithms in a model to minimize the consequences of any one failing, may not yield truly robust systems because they do not resolve the underlying problem.

To be clear, it is too early to draw definitive conclusions. The point is that applying machine learning to cybersecurity presents a set of intertwined challenges. At a minimum, defenders will have to ensure their systems keep up with constantly evolving threats. But the same capabilities that enable

---

* Sven Krasser, chief scientist and vice president of CrowdStrike, observes that even with a detection system with a 99 percent success rate, an attacker can defeat it with over a 99 percent chance of success with 500 tries. See National Academies of Sciences, Engineering, and Medicine, *Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop* (Washington, DC: The National Academies Press, 2019), page 43.

systems to adapt may put them at risk of being "mistrained" in ways that leave them vulnerable to targeted attacks. And if it is possible that there are inherent limits on robustness, defenders could be forced to make tradeoffs between different threats.

The defender's predicament

Machine learning may solve some long-standing problems for defenders while creating new ones. In many contexts, defenses sufficient to deal with that vast majority of malicious threats will be good enough. States, however, need to ensure the viability of defenses against not just general malicious activity, but specific pacing threats (e.g. China or Russia in the United States' case). The possibility of an adversary exploiting a hidden failure mode in a defense may become an acute concern. Yet states may have limited options for ensuring the robustness of defenses, each of which may necessitate intruding into their adversaries' (or third parties') networks before an attack occurs:

(1) Overcoming the limitations of training, testing and verification: Generally speaking, knowledge of adversaries' capabilities enables proper threat modeling and hardening of defenses. Machine learning could amplify the benefits of insights into the evolving threat landscape—and the potential costs of falling behind the latest trends. The better the training data on attacks are, the better the defensive model against those attacks will be. Historical data will diminish in value as adversaries change tactics and the landscape shifts, creating a constant incentive to continually gather information on evolving adversary tactics. Moreover, these incentives could be even stronger if there are inherent limits on the robustness of machine learning defenses. The defender may have to choose a subset of potential attacks to prioritize when training a defense within a vast range of possible attacks.[46] Verifying the system's robustness against a specific adversary might depend on anticipating their likely attack methodology. Intruding into the adversary's networks (or a third-party network that adversary may be operating inside) to gain advanced warning of their capabilities could thus guide the defender's efforts and make this problem far more tractable.

(2) Enabling countermeasures to a specific adversary's attacks: A defender can painstakingly try to harden a defense against the vast range of possible attacks. But a much simpler option may exist: peer into the attacker's own networks to gain the information necessary to mitigate an attack through traditional cyber defense. This could include discovering and patching a software vulnerability used by the attacker or creating a signature of malware in order to detect it, essentially "inoculating" the defense. This would have the

added benefit of scalability; a defender could inoculate defenses deployed in a range of settings rather than having to orchestrate their retraining.* Rapidly inoculating defenses might be especially necessary in a period of heightened tensions when an attack by an adversary is anticipated.

(3) Leapfrogging the innovations of others: Unlike experimental settings that typically feature one attacker and one defender, cyberspace features many actors who learn from and appropriate others' tools and techniques. With cybersecurity in general, a state can expect its adversaries to adapt and improve their capabilities against other states' defenses. The fact that attacks tend to transfer from one machine learning model to another suggests that observing successful attacks against another's defenses can yield specific, valuable information on how to improve one's own. A state might even probe another actor's defenses to try and extract the model and copy it for its own defense.

* U.S. Cyber Command's "malware inoculation initiative," which publishes information discovered on adversaries' capabilities to improve private sector defenses, demonstrates the potential scalability of this approach. Erica Borghard and Shawn Lonergan, "U.S. Cyber Command's Malware Inoculation: Linking Offense and Defense in Cyberspace," Net Politics, April 22, 2020, https://www.cfr.org/blog/us-cyber-commands-malware-inoculation-linking-offense-and-defense-cyberspace.

Artificial Intelligence and Cyber Stability

Artificial intelligence could transform cyber operations at a time when cyber competition among states is heating up. This analysis has focused on the potential operational imperatives machine learning could create, but these operations would not play out in a vacuum. They would occur within this strategic context, in which states may be both "attackers" and "defenders" in a constant struggle for advantage. The stakes are no less than protecting core national interests and potentially crucial military advantages in a conflict. Cyber competition may drive states to hack machine learning defenses. Could machine learning, in turn, destabilize cyber competition?

The escalation dynamics of cyber engagements remain a subject of contention. Real-world cyber operations have rarely provoked forceful responses.[47] This has led some scholars to propose that inherent characteristics of cyber capabilities or cyber competition limit the potential for escalation. Others are less sanguine. Jason Healey and Robert Jervis argue that cyber competition has steadily intensified as the scope and scale of cyber operations have expanded over three decades.[48] The forces containing this competition to manageable thresholds may not hold indefinitely. Moreover, they argue that even if cyber operations can be stabilizing in some circumstances, in a crisis their characteristics could accelerate the path to conflict.

Cyber competition already has the ingredients needed for escalation to real-world violence, even if these ingredients have yet to come together in the right conditions. The aim here is simply to show how machine learning could potentially amplify these risks. This follows two of the potential escalation pathways Healey and Jervis identify. The first concerns the factors fueling the steady intensification of cyber competition, which could eventually cross a threshold triggering a crisis. The second concerns the characteristics of cyber operations that may pressure states to launch attacks in a crisis.

(1) Machine learning could fuel the intensification of cyber competition.

Even as states' cyber operations have become more aggressive in some respects, they have largely remained well below the threshold likely to trigger retaliation. The vast majority consist of acts of espionage and subversion in the "gray zone" between war and peace. Some attribute this apparent stability to dynamics governing cyber competition below the use of force that are inherently self-limiting.[49] But Healey and Jervis argue that this stability may be tenuous. In some conditions, cyber competition leads to "negative feedback loops" that diffuse tensions. In others, it can lead to "positive

feedback loops," whereby cyber operations by one state incite operations by another.[50] Positive feedback can occur when cyber operations generate fears of insecurity. A state may intrude into another's networks simply to maintain situational awareness or to secure its own networks against the target's offensive capabilities. But because the same intrusion for espionage could pave the way to launch an attack, the target of the intrusion may view this as offensive and respond by engaging in their own counter intrusions.*

How might machine learning change these dynamics? The above analysis of offensive and defensive imperatives suggests the potential to amplify positive feedback loops in three ways:

First, machine learning may increase the perceived salience of informational advantages over an adversary and the fear of falling behind. Offensive operations targeting machine learning attack vectors may have to be tailored to the precise defensive configuration.† Defending against such attacks may require the ability to anticipate the particular deception created by the attacker. The resulting strategic dynamic may resemble the game of poker: Your best move depends on what your opponent has in their hand. Whatever can be done in advance to figure out the opponent's hand—or "stack the deck"—may prove tempting.

Second, machine learning may incentivize states to conduct intrusions into adversaries' networks even earlier in anticipation of future threats. Whether attacking machine learning systems or defending against such attacks, the options with the greatest chance of success may also require the earliest action. Reaching an isolated target may necessitate sabotaging a machine learning defense before it is deployed if a black-box attack would be infeasible. Similarly, hardening a defense against an attack may require gaining information on an attacker's capabilities well before they are launched. States tend to hedge against uncertainties. They may be forced to make decisions to take action in the present based on possible future

---

* This dynamic, whereby one state's actions to secure itself create fear in another, raising the potential for misinterpretation and escalation, is similar to the political science concept of the security dilemma. For an overview of the security dilemma and its application to cybersecurity, see Robert Jervis, "Cooperation Under the Security Dilemma," *World Politics* 30, no. 2 (1978): 167–214; Ben Buchanan, *The Cybersecurity Dilemma* (New York, NY: Oxford University Press, 2016).

† Notably, in their effort to defeat proposed defenses against adversarial examples, Tramer et al. found that "no single strategy would have been sufficient for all defenses. This underlines the crucial fact that adaptive attacks cannot be automated and always require appropriate tuning to a given defense." Tramer et al., "On Adaptive Attacks to Adversarial Example Defenses."

offensive or defensive needs. The result may be to lower the threshold of perceived threat sufficient to motivate such action.

Third, machine learning may further blur the line between offensive and defensive cyber operations. If merely interacting with a defensive system could extract information needed to engineer an attack to defeat it, states may be prone to view any interaction as possible preparation for an attack. Similarly, a state may gain access to a training environment to copy a defensive model, but the target may fear the model has been reverse-engineered and fatally compromised, enabling an attack.

In short, states may perceive that the stakes of gaining an edge over adversaries are rising, requiring even more proactive efforts in anticipation of future needs, while simultaneously making the same efforts by adversaries seem even more threatening. In the right conditions, positive feedback loops may become more likely to cause an engagement to cross a threshold triggering a crisis. More predictably, these dynamics might motivate risky or destabilizing cyber operations by states—particularly those seeking asymmetric advantages and willing to tolerate collateral damage. Several concerning scenarios stand out:

- Systemic compromises: Contractors or open source projects may present opportunities to insert backdoors into models that make their way into harder to reach targets. The danger of such operations is that a systemic compromise could leave a wide swath of civilian and governmental applications vulnerable. Malware designed to exploit the backdoor could inadvertently propagate to other systems. As with any backdoor inserted into a product, there is no guarantee another malicious actor could not discover and exploit it.
- Poisoning the waters: A cruder tactic than inserting a backdoor would simply be an indiscriminate attempt to degrade cybersecurity applications. An attacker with little regard for collateral damage might flood a malware repository with tainted samples designed to mistrain machine learning systems relying on the data.
- Reckless operations: States may be tempted to accept certain operational risks to circumvent machine learning defenses. For instance, an attacker may employ capabilities with greater autonomy to avoid reliance on external command and control servers, which would risk detection.[51] Absent human control, such capabilities might carry greater risk of unintended impacts that spread beyond the target network. An attacker might also sabotage a defense to create an offensive option that unintentionally exposes the targeted network to other attackers. Sabotaging the systems that protect an adversary's

critical infrastructure, for instance, might backfire catastrophically if it creates an opportunity for a third party to launch an attack and trigger a crisis.

- Attacks on trust: An attacker might not need to break a machine learning defense if they can undermine the defender's confidence in it. A case alleged against the cybersecurity vendor Kaspersky illustrates the possibility. In 2015, the company was accused of uploading fake malware samples to VirusTotal, an open source service that aggregates information from cybersecurity vendors to improve collective defenses. The fake samples were designed to cause competing antivirus systems to flag legitimate files, creating problems for clients and potentially hurting their brands.[52] Manipulating a machine learning system to trigger false positives could similarly undermine confidence in the model.

(2) Machine learning could exacerbate the characteristics of cyber operations that undermine crisis stability.

In some cases, cyber operations might help avoid a crisis by diffusing tensions.* However, if a crisis breaks out, cyber capabilities create pressures that could accelerate the path to conflict. Healey and Jervis note the widespread perception that cyber capabilities have maximal effect when the attacker has the benefits of surprise and initiative.[53] If conflict appears imminent, such first-mover advantages might tempt states to launch preemptive cyberattacks against command, control, and communications capabilities to degrade or disable an adversary's military forces. Short of actually launching an attack, states would have strong incentives to begin preparations to do so by intruding into their opponent's networks.

The inherent ambiguity of cyber intrusions creates a recipe for misperception in such a context. Intrusions for espionage purposes may appear indistinguishable from those laying the groundwork for attacks, or "operational preparation of the environment" (OPE). As Buchanan and Cunningham argue, this creates the potential for escalation resulting from a

---

* Cyber operations could act as a "pressure valve" by creating options to respond to provocations that are potentially less escalatory than kinetic force both in their direct impacts and impacts on perceptions. See Benjamin Jensen and Brandon Valeriano, "What do We Know About Cyber Escalation? Observations from Simulations and Surveys" (Atlantic Council, November 2019), https://www.atlanticcouncil.org/wp-content/uploads/2019/11/What_do_we_know_about_cyber_escalation_.pdf; Sarah Kreps and Jacquelyn Schneider, "Escalation Firebreaks in the Cyber, Conventional, and Nuclear Domains: Moving Beyond Effects-Based Logics," *Journal of Cybersecurity* 5, no. 1 (January 1, 2019): 1–11, https://doi.org/10.1093/cybsec/tyz007.

miscalculated response to an intrusion.[54] One side might discover an intrusion in a crisis—even one that occurred months before the crisis began—and, misinterpreting it as an imminent attack, may face sudden pressure to launch a counterattack.

Here again, the potential offensive and defensive imperatives created by machine learning could exacerbate these risks:

First, states may feel even greater pressure to gain advantages through intrusions early in a crisis. The time needed to engineer an attack under black-box conditions, or retrain a defense to ensure robustness against a possible imminent attack by an adversary, may translate to increased pressure to try and quickly gain information on an adversary's capabilities if the state does not already possess it.*

Second, the indistinguishability of espionage from OPE may be even more problematic. A state that detects an intrusion or a compromised training process might have no way to rapidly discern whether the attacker has discovered a flaw that would defeat the system or to evaluate the robustness of the defense. If defenses are believed to be fragile in the face of novel attacks this could become an acute source of anxiety. Faced with fewer options to rule out worst-case scenarios, the state may be more likely to escalate in response.

Third, machine learning could create additional sources of uncertainty that induce potential "use it or lose it" dilemmas. Changes in the target environment can already throw off meticulously-planned offensive operations. The shelf-life of an offensive operation might be even shorter if it must be tailored to the precise configuration of machine learning defenses that could evolve over time. If a state has prepared such capabilities, the temptation in a crisis may be to use them rather than risk them becoming obsolete. The sudden discovery of a critical flaw in a defensive machine learning system with no easy remedy might similarly force the defender to contemplate whether to preempt a possible attack.

The threat to crisis stability arises from this unique combination of uncertainties and anxieties at the technical and strategic levels. Machine learning seems

---

* If conducting black-box attacks on machine learning systems proves time-consuming, this might actually be stabilizing in some circumstances: as Borghard and Lonergan argue, the time needed to develop offensive options makes them a less effective tool of escalation in response to an attack–a state cannot simply conjure up cyber options for immediate retaliation. See Erica Borghard and Shawn Lonergan, "Cyber Operations as Imperfect Tools of Escalation," *Strategic Studies Quarterly* (Fall 2019): 122-145.

capable of compounding these and, in the heat of a crisis, increasing the potential for serious misperception and miscalculation.

Mitigating scenarios

As stated at the outset, this study explores a possible worst-case scenario for the future of AI-cyber capabilities. The threat to stability stems from the potential for machine learning to create offensive and defensive imperatives that incentivize states to intrude into their adversaries' networks. But it is worth briefly revisiting the possibility that machine learning could evolve in ways that fundamentally change these imperatives. Describing the current state of the art, Bruce Schneier compares machine learning security to the field of cryptography in the 1990s: "Attacks come easy, and defensive techniques are regularly broken soon after they're made public."[55] The field of cryptography, however, matured and encryption is now one of the strongest aspects of cybersecurity. Eventually, a more mature science of machine learning security may likewise yield systems highly robust without the constant threat of becoming vulnerable to targeted attacks.

However, as this relates to the incentives to intrude into adversaries' networks, it only solves the defensive side of the equation. A machine learning defense could be robust to an adversary's attack even without advanced knowledge of their capabilities. On the other hand, attackers may face even greater incentives to intrude into target networks early and aggressively. If attackers cannot count on discovering ways to defeat a system once it is deployed, sabotaging its development or compromising its supply chain may be seen as even more necessary offensive options.

Alternatively, machine learning security may hit a dead end. Systems may remain fundamentally vulnerable in dynamic, adversarial conditions. In such a scenario, cybersecurity would in all likelihood still benefit from machine learning applications as it does now, but not in ways that fundamentally change the cat-and-mouse game. In this case, offensive operations may not depend on early intrusions any more than in the status quo; attackers would likely be able to find ways to defeat defenses that do not depend on compromising them well in advance. Defenders, however, might face stronger pressure to intrude into adversaries' networks to try and harden potentially fragile systems against their capabilities. The situation for defenders could become untenable if attackers benefit from *offensive* applications of machine learning.[56]

The point of this cursory analysis is to show that even if the broad trajectory of machine learning changes, the potential for destabilization may remain. In any event, the present challenges facing machine learning do not appear

likely to be resolved soon. To Schneier, machine learning security is at the level of maturity of cryptography in the 1990s, but Nicholas Carlini, a leading expert on machine learning security, paints an even bleaker picture. In a November 2019 presentation, he compared machine learning security to cryptography in the 1920s, suggesting that the field has not even developed the right metrics to gauge progress toward solving these fundamental problems.[57]

Implications for policy

Efforts are underway to understand and address the threats to machine learning systems.[58] A key takeaway from this study is that deploying machine learning for cybersecurity presents a unique set of challenges arising from the interaction of technical characteristics and strategic imperatives. These challenges must be addressed not only via technical solutions but at the level of policy and strategy. Even with the "known-unknowns," several imperatives emerge from this forecast:

- First, machine learning may present inexorable tradeoffs for cyber defense. Machine learning defenses may mitigate some threats while introducing new attack vectors. And the ability to adapt to evolving threats may put systems at constant risk of becoming vulnerable. Decision-makers need basic tools to inform risk-based decisions about when and how to employ such systems. This includes frameworks and metrics to evaluate systems deployed in crucial contexts: e.g., diagnosability or auditability, resilience to poisoning or manipulation, and the ability to "fail gracefully" (meaning a model's failure does not cause catastrophic harm to functions dependent upon it).[59] Decisions and policies, such as those regarding the disclosure of machine learning vulnerabilities or the publication of offensive security research that might enable attackers, will also have to be adapted to the unique characteristics of machine learning.*
- Second, secure deployment of machine learning depends on guarding against attempts by adversaries to broadly compromise or degrade the development process. Attacks will not always be direct; adversaries may exploit trust in common services, like the aforementioned case involving VirusTotal. They may further blur the

* Machine learning security would benefit, for instance, from standards analogous to the Common Vulnerability Scoring System, used to evaluate the severity of software vulnerabilities and inform decisions about patching. "Common Vulnerability Scoring System Version 3.1: User Guide," Forum of Incident Response and Security Teams, July 2019, https://www.first.org/cvss/user-guide.

lines between threats such as industrial espionage and strategic sabotage. Given the premium on "box knowledge," threats to the confidentiality of public or private data and algorithms should be treated as threats to the integrity of applications. Securing machine learning demands collective efforts by the government and private sector to secure the supply chain, open source development projects, data repositories, and other critical inputs.[60]

- Third, managing tension and avoiding escalation in the conduct of cyber espionage and offensive operations will become even more important—especially as the imperative to gain information on adversaries' offensive capabilities and their own machine learning defenses increases. Operators need to understand the potential impacts of operations against machine learning in sensitive contexts, and will need to understand how adversaries will perceive their operations. If machine learning could amplify positive feedback loops it is worth examining the broader implications for U.S. cyber strategy, which is premised on the stabilizing effects of "persistent engagement" with adversaries.[61] Communication with adversaries to clarify strategic intentions will help avoid misinterpretation. Further, now is the time to explore forms of mutual restraint regarding the most destabilizing offensive activities targeting machine learning.

Conclusion

The pressure to harness artificial intelligence to deal with evolving offensive cyber capabilities will only grow. Precisely because machine learning holds both promise and peril for cybersecurity, a healthy dose of caution is needed when embracing these capabilities. Decisions made now with respect to the development and adoption of machine learning could have far-reaching consequences for security and stability. Decision-makers must avoid having to relearn lessons from cybersecurity in general, including the pitfalls of over-reliance on defenses at the expense of a more holistic approach to risk management. The stakes of securing machine learning will rise as it is incorporated into a wide range of functions crucial to economic and national security. The incentives to gather intelligence or even sabotage the development of defensive systems might weigh just as heavily with other strategic areas of application. If this competition is not managed, states may head down a path destructive for all.

Acknowledgments

The author would like to thank Jason Healey, Paul Scharre, Trey Herr, Jon Bateman, John Bansemer, Ben Buchanan, and Andrew Lohn for their invaluable feedback and comments on earlier versions of this manuscript.

Endnotes

[1] Mariarosaria Taddeo and Luciano Floridi, "Regulate Artificial Intelligence to Avert Cyber Arms Race," *Nature* 556, no. 7701 (April 2018): 296–98.

[2] Elsa B. Kania, "'AI Weapons' in China's Military Innovation" (Brookings Institution, April 2020), https://www.brookings.edu/research/ai-weapons-in-chinas-military-innovation/; Rod Thornton and Marina Miron, "Towards the 'Third Revolution in Military Affairs,'" *The RUSI Journal* 165, no. 3 (April 15, 2020): 12–21.

[3] Zachary Fryer-Biggs, "Twilight of the Human Hacker," The Center for Public Integrity, Sept. 13, 2020, https://publicintegrity.org/national-security/future-of-warfare/scary-fast/twilight-of-the-human-hacker-cyberwarfare/.

[4] One recent estimate valued the total market for AI-related cybersecurity services at $8.8 billion, projected to grow to $38.2 billion by 2026. "Artificial Intelligence in Cybersecurity Market," MarketsandMarkets, May 2019, https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-security-market-220634996.html.

[5] For an overview of machine learning, see Ben Buchanan, "The AI Triad and What It Means for National Security Strategy" (Center for Security and Emerging Technology, August 2020), https://live-cset-georgetown.pantheonsite.io/research/the-ai-triad-and-what-it-means-for-national-security-strategy/.

[6] Sean Park et al. "Generative Malware Outbreak Detection," (Trend Micro Research, 2019), https://documents.trendmicro.com/assets/white_papers/GenerativeMalwareOutbreakDetection.pdf.

[7] Ahmad Ridley, "Machine Learning for Autonomous Cyber Defense," *The Next Wave* 22, no. 1 (2018): 7–14.

[8] See comments by Sven Krasser in National Academies of Sciences, Engineering, and Medicine, *Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop* (Washington, DC: The National Academies Press, 2019).

[9] "Machine Learning in the Age of Cyber AI," Darktrace, 2019. https://www.darktrace.com/en/resources/wp-machine-learning.pdf.

[10] Micah Musser and Ashton Garriott, "Machine Learning and Cybersecurity: Hype and Reality" (Center for Security and Emerging Technology, forthcoming).

[11] CrowdStrike, for instance, employs a dedicated team of human experts to maintain oversight of its machine learning-based tools and ensure data is collected and analyzed properly. See comments by Sven Krasser in National Academies of Sciences, Engineering, and Medicine, *Implications of Artificial Intelligence for Cybersecurity*.

[12] David Silver et al., "Mastering the Game of Go without Human Knowledge," *Nature* 550, no. 7676 (October 18, 2017): 354–59.

[13] Joe Slowik "The Myth of the Adversary Advantage," Dragos, June 19, 2018, https://www.dragos.com/blog/industry-news/the-myth-of-the-adversary-advantage/.

[14] Richard Colbaugh and Kristin Glass, "Predictive Moving Target Defense," (Sandia National Laboratories, 2012), https://www.osti.gov/servlets/purl/1117315.

[15] Ridley, "Machine Learning for Autonomous Cyber Defense"; Thanh Thi Nguyen and Vijay Janapa Reddi, "Deep Reinforcement Learning for Cyber Security," *arXiv [cs.CR]* (June 13, 2019), arXiv, http://arxiv.org/abs/1906.05799; Seamus Dowling, Michael Schukat, and Enda Barrett, "Improving Adaptive Honeypot Functionality with Efficient Reinforcement Learning Parameters for Automated Malware," *Journal of Cyber Security Technology* 2, no. 2 (April 3, 2018): 75–91.

[16] Taha Eghtesad, Yevgeniy Vorobeychik, and Aron Laszka, "Deep Reinforcement Learning Based Adaptive Moving Target Defense," *arXiv [cs.CR]* (November 27, 2019), arXiv, http://arxiv.org/abs/1911.11972.

[17] Ram Shankar Siva Kumar et al., "Failure Modes in Machine Learning Systems," *arXiv [cs.LG]* (November 25, 2019), arXiv, http://arxiv.org/abs/1911.11034.

[18] Anish Athalye et al., "Synthesizing Robust Adversarial Examples," *arXiv [cs.CV]* (July 24, 2017), arXiv, http://arxiv.org/abs/1707.07397.

[19] Sandy Huang et al., "Adversarial Attacks on Neural Network Policies," *arXiv [cs.LG]* (February 8, 2017), arXiv, http://arxiv.org/abs/1702.02284; Adam Gleave et al., "Adversarial Policies: Attacking Deep Reinforcement Learning," *arXiv [cs.LG]* (May 25, 2019), arXiv, http://arxiv.org/abs/1905.10615.

[20] Florian Tramèr et al., "The Space of Transferable Adversarial Examples," *arXiv [stat.ML]* (April 11, 2017), arXiv, http://arxiv.org/abs/1704.03453.

[21] Geoffrey I. Webb et al., "Characterizing Concept Drift," *Data Mining and Knowledge Discovery* 30, no. 4 (July 1, 2016): 964–94.

[22] Tegjyot Singh Sethi et al., "A Dynamic-Adversarial Mining Approach to the Security of Machine Learning," *arXiv [cs.LG]* (March 24, 2018), arXiv, http://arxiv.org/abs/1803.09162.

[23] Myriam Abramson, "Toward Adversarial Online Learning and the Science of Deceptive Machines," in *2015 AAAI Fall Symposium Series*, 2015, https://www.aaai.org/ocs/index.php/FSS/FSS15/paper/viewPaper/11661.

[24] J. Kelly et al., "Adversarially Adapting Deceptive Views and Reconnaissance Scans on a Software Defined Network," in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2019, 49–54.

[25] Lerrel Pinto et al., "Robust Adversarial Reinforcement Learning," *arXiv [cs.LG]* (March 8, 2017), arXiv, http://arxiv.org/abs/1703.02702.

[26] Una-May O'Reilly et al., "Adversarial Genetic Programming for Cyber Security: A Rising Application Domain Where GP Matters," *arXiv [cs.CR]* (April 7, 2020), arXiv, http://arxiv.org/abs/2004.04647.

[27] "Cylance, I Kill You!," Skylight Cyber, July 18, 2019, https://skylightcyber.com/2019/07/18/cylance-i-kill-you/.

[28] "Cylance, I Kill You!"

[29] Anirban Chakraborty et al., "Adversarial Attacks and Defences: A Survey," *arXiv [cs.LG]* (September 28, 2018), arXiv, https://arxiv.org/abs/1810.00069.

[30] Giorgio Severi et al., "Exploring Backdoor Poisoning Attacks Against Malware Classifiers," *arXiv [cs.CR]* (March 2, 2020), arXiv, http://arxiv.org/abs/2003.01031.

[31] Gleave et al., "Adversarial Policies: Attacking Deep Reinforcement Learning"; Yansong Gao et al., "Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review," *arXiv [cs.CR]* (July 21, 2020), arXiv, http://arxiv.org/abs/2007.10760.

[32] Yi Han et al., "Reinforcement Learning for Autonomous Defence in Software-Defined Networking," *arXiv [cs.CR]* (August 17, 2018), arXiv, http://arxiv.org/abs/1808.05770.

[33] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot, "Making Machine Learning Robust Against Adversarial Inputs," *Communications of the ACM* 61, no. 7 (June 2018): 56–66.

[34] Florian Tramer et al., "On Adaptive Attacks to Adversarial Example Defenses," *arXiv [cs.LG]* (February 19, 2020), arXiv, http://arxiv.org/abs/2002.08347.

[35] Sethi et al., "A Dynamic-Adversarial Mining Approach to the Security of Machine Learning."

[36] Alex Kantchelian et al., "Approaches to Adversarial Drift," in *Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security*, AISec '13 (New York, NY: Association for Computing Machinery, 2013), 99–110.

[37] Sven Herpig, "Securing Artificial Intelligence: Part 1: The attack surface of machine learning and its implications" (Stiftung Neue Verantwortung, October 2019), https://www.stiftung-nv.de/sites/default/files/securing_artificial_intelligence.pdf.

[38] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," *arXiv [cs.CR]* (August 22, 2017), arXiv, http://arxiv.org/abs/1708.06733; Yuanshun Yao et al., "Latent Backdoor Attacks on Deep Neural Networks," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19 (New York, NY: Association for Computing Machinery, 2019), 2041–55.

[39] See, for instance, Nicholas Carlini, "Are Adversarial Example Defenses Improving?," February 20, 2020, https://nicholas.carlini.com/writing/2020/are-adversarial-exampe-defenses-improving.html.

[40] Ian Goodfellow and Nicolas Papernot, "Is Attacking Machine Learning Easier than Defending It?," *Cleverhans Blog*, February 15, 2017, http://www.cleverhans.io/security/privacy/ml/2017/02/15/why-attacking-machine-learning-is-easier-than-defending-it.html.

[41] Tramer et al., "On Adaptive Attacks to Adversarial Example Defenses."

[42] Trapit Bansal et al., "Emergent Complexity via Multi-Agent Competition," *arXiv [cs.AI]* (October 10, 2017), arXiv, http://arxiv.org/abs/1710.03748.

[43] David Brumley, "Why I'm Not Sold on Machine Learning in Autonomous Security," CSO, August 27, 2019, https://www.csoonline.com/article/3434081/why-im-not-sold-on-machine-learning-in-autonomous-security.html.

[44] Nicholas Carlini et al., "On Evaluating Adversarial Robustness," *arXiv [cs.LG]* (February 18, 2019), arXiv, http://arxiv.org/abs/1902.06705.

[45] Richard Colbaugh and Kristin Glass, "Asymmetry in Coevolving Adversarial Systems," in *2016 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 2016, 360–67.

[46] Justin Gilmer et al., "Motivating the Rules of the Game for Adversarial Example Research," *arXiv [cs.LG]* (July 18, 2018), arXiv, http://arxiv.org/abs/1807.06732.

[47] Brandon Valeriano, Benjamin Jensen, and Ryan Maness, *Cyber Strategy: The Evolving Character of Power and Coercion* (New York: Oxford University Press, 2018).

[48] Jason Healey and Robert Jervis, "The Escalation Inversion and Other Oddities of Situational Cyber Stability," *Texas National Security Review* 3, no. 4 (2020), https://tnsr.org/2020/09/the-escalation-inversion-and-other-oddities-of-situational-cyber-stability/#_ftn28.

[49] See, for instance, Joshua Rovner, "Cyber War as an Intelligence Contest," War on the Rocks, September 16, 2019, https://warontherocks.com/2019/09/cyber-war-as-an-intelligence-contest/; Michael P. Fischerkeller and Richard P. Harknett, "What Is Agreed Competition in Cyberspace?," *Lawfare*, February 19, 2019, https://www.lawfareblog.com/what-agreed-competition-cyberspace.

[50] Healey and Jervis, "The Escalation Inversion."

[51] Ben Buchanan et al. "Automating Cyber Attacks: Hype and Reality" (Center for Security and Emerging Technology, November 2020), https://cset.georgetown.edu/research/automating-cyber-attacks/.

[52] Joseph Menn, "Exclusive: Russian Antivirus Firm Faked Malware to Harm Rivals - Ex-Employees," Reuters, August 4, 2015, https://www.reuters.com/article/us-kaspersky-rivals/exclusive-russian-antivirus-firm-faked-malware-to-harm-rivals-ex-employees-idUSKCN0QJ1CR20150814.

[53] Healey and Jervis, "The Escalation Inversion."

[54] Ben Buchanan and Fiona S. Cunningham, "Preparing the Cyber Battlefield: Assessing a Novel Escalation Risk in a Sino-American Crisis," *Texas National Security Review* 3, no. 4 (2020), https://tnsr.org/2020/10/preparing-the-cyber-battlefield-assessing-a-novel-escalation-risk-in-a-sino-american-crisis/.

[55] Bruce Schneier, "Attacking Machine Learning Systems," *Computer* 53, no. 5 (May 2020): 78–80.

[56] Buchanan et al. "Automating Cyber Attacks."

[57] Nicholas Carlini, "On Evaluating Adversarial Robustness" (2019 Conference on Applied Machine Learning in Information Security, Washington, DC, October 26, 2019), https://www.camlis.org/2019/keynotes/carlini.

[58] See, for instance, the Adversarial Machine Learning Threat Matrix jointly developed by 12 major industry and academic organizations. Ram Shankar Siva Kumar and Ann Johnson, "Cyberattacks against machine learning systems are more common than you think," Microsoft, October 22, 2020, https://www.microsoft.com/security/blog/2020/10/22/cyberattacks-against-machine-learning-systems-are-more-common-than-you-think/.

[59] Jacob Steinhardt and Helen Toner, "Why Robustness Is Key to Deploying AI" (Brookings Institution, June 8, 2020), https://www.brookings.edu/techstream/why-robustness-is-key-to-deploying-ai/.

[60] For an overview of the supply chain for machine learning see Sven Herpig, "Understanding the Security Implications of the Machine-Learning Supply Chain: Securing Artificial Intelligence – Part 2" (Stiftung Neue Verantwortung, October 2020), https://www.stiftung-nv.de/sites/default/files/understanding_the_security_of_the_machine-learning_supply_chain.pdf.

[61] Jason Healey, "The implications of persistent (and permanent) engagement in cyberspace," *Journal of Cybersecurity* 5, no. 1 (2019): 1-15.