December 2021

Making AI Work for Cyber Defense

The Accuracy-Robustness Tradeoff

CSET Issue Brief

CSET
CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHOR
Wyatt Hoffman

Executive Summary

Cyber threats are multiplying and escalating. AI could exacerbate the problem or be part of the solution. Innovations in machine learning (ML) methodologies have already proven their usefulness for cybersecurity. But can ML-enabled defenses deployed at scale contend with adaptive attackers? To level the playing field for defenders, machine learning must be able to perform reliably under sustained pressure from offensive campaigns—without constant human supervision.

Yet machine learning carries with it new security challenges. ML systems rely on patterns in data to make predictions, which can be manipulated by attackers to evade defenses. Techniques to make ML systems more robust to these deceptions often harm their accuracy. They essentially prevent the system from looking for certain patterns useful for making predictions, and thus make it more prone to error. This accuracy-robustness tradeoff creates a problem, for example, for an ML-based antivirus system continuously adapting to keep pace with evolving malware. The developer can carefully supervise the system and harden it to deceptive attacks, but doing so impedes its ability to accurately detect new malware. Deploying ML systems to deal with dynamic threats will require constantly balancing these risks.

This balancing act will present a serious challenge for automating ML-cyber defenses at scale. Defenders have options to manage risk tradeoffs by employing multiple systems, hardening them against specific threats, and offsetting their limitations with non-ML tools. But defenders' reliance on machine learning will change the threat landscape in ways that undermine their ability to proactively make these tactical choices. Machine learning will expand the attack surface by creating new dependencies on widely relied upon tools, services, and data sources. Persistent attackers will exploit the compromises defenders make between accuracy and robustness. The most sophisticated adversaries may use ML capabilities themselves to counter ML-based defenses. Defenders need tools to gauge and improve accuracy and robustness that can cope with these innovations in adversary tactics and offensive campaigns.

The imperative for policy and strategy will be to put defenders on solid ground—enabling and empowering them to manage these tradeoffs at a tactical level. This includes four recommendations for government efforts to shape the trajectory of the emerging ML-cybersecurity ecosystem toward a more tenable situation for defenders:

(1) Build security into the process of ML design and development. The typical approach to development that prioritizes efficiency and accuracy will not suffice. Cybersecurity calls for systems that can learn continuously while under pressure from attackers and demonstrate provable robustness to real-world threats.

(2) Promote resilience through system diversity and redundancy. Machine learning will never be foolproof. Decisionmakers must set thresholds for critical security settings that limit the impacts of ML vulnerabilities, including offsetting risks with ML and non-ML tools and safeguards.

(3) Manage the risk that cuts across the ML and cybersecurity ecosystem. This requires mapping out and protecting critical dependencies, from reliance on shared datasets and open-source tools to feedback loops in deployed systems.

(4) Counter strategic rivals' attempts to compromise and sabotage ML development. Rivals will try to infiltrate development processes, whether to extract information in order to reverse engineer ML systems or to manipulate training data to corrupt them. Successful defense at a tactical level will depend on thwarting offensive campaigns that aim to fatally compromise ML defenses before they are even deployed.

Introduction

Cyber defenses struggle to keep up with ever more sophisticated offensive capabilities. Artificial intelligence—specifically cutting-edge ML methodologies—has already begun to ease the burden on defenders. But the offense constantly evolves, and ML capabilities will add to top-tier adversaries' already potent toolkit. Whether machine learning gives defenders an edge or compounds their problems depends on the ability of automated ML-cyber defenses to contend with these escalating threats. As the National Security Commission on Artificial Intelligence (NSCAI) warned, "Defending against AI-capable adversaries operating at machine speeds without employing AI is an invitation to disaster."[1]

However, evidence of pervasive vulnerabilities in ML systems raises serious questions about their reliability under sustained pressure by adaptive attackers. Machine learning looks for patterns in data to develop a model used to make predictions. This methodology can be extremely effective, but it makes ML systems susceptible to error and malicious interference. An attacker able to manipulate data inputs could create a deceptive pattern to trick the model. These vulnerabilities are widely acknowledged, including in national security circles.[2] A growing field of study aims to develop robust ML systems that are secure against deceptive attacks.

The problem is that measures to improve robustness to deception often degrade the overall accuracy of ML systems' predictions. This has been observed in a range of settings, giving reason to think that accuracy and robustness may be "fundamentally at odds."[3] This tradeoff creates a tension between two core objectives of cyber defense: defenders need ML systems to accurately identify both generic cyberattacks and attacks specifically targeting the ML model. But these are two distinct problems, and the ML system can't be optimized for both at the same time. In fact, solving one might make the other worse. For example, an ML-based antivirus system may be highly accurate, identifying 99 percent of malware, but easily deceived by malware "camouflaged" to trick the ML model. Improving robustness might eliminate this attack vector, but make the system less accurate at identifying malware in general.

Defenders have options to manage the tradeoffs and limitations of ML systems. They can strike a balance between enabling the system to adapt to accurately detect new threats and hardening it against deception. Many cybersecurity services already rely on ML systems under close supervision and in concert with other techniques that limit the consequences of a system's failure.

Whether this approach can scale, however, is another question. Managing vulnerabilities in ML systems will be increasingly difficult as they create new dependencies and expand the attack surface. Attackers will change tactics and strategies to exploit the compromises defenders make between accuracy and robustness. The most sophisticated and determined adversaries will have access to tools, including offensive ML capabilities, to counter ML defenses. Any approach to deploying ML-cyber defenses must be viable against evolving offensive campaigns, not just one-off attacks.

This report focuses on these challenges that will emerge and grow with reliance on automated ML-cyber defenses. It argues that understanding the technical tradeoffs and limitations of machine learning will be necessary to anticipate how its deployment at scale will shape cyber offense and defense. This is not the most pressing issue for researchers and cybersecurity vendors preoccupied with the immediate problems of acquiring sufficient training data, computing resources, and other technical and practical hurdles. But decisions made even at this early stage—from the fundamental design of systems to their implementation across users' networks—may be crucial in either reducing or exacerbating the challenge of deploying machine learning securely.

Even those skeptical of the prospects of machine learning for cybersecurity should consider the implications of this accuracy-robustness tradeoff. While a major problem for defense, it will not inhibit the offensive use of machine learning—attackers often simply need a tool that works once, not one that performs reliably and consistently under changing conditions.

This report begins by laying out the stakes: what machine learning has to offer cyber defense. Drawing from recent research on ML security, it then explores the accuracy-robustness tradeoff and the challenges it will create for deploying automated ML-cyber defenses at scale. Finally, it identifies four key areas in which policymakers can shape the trajectory of the ML and cybersecurity ecosystem to enable defenders to manage the tradeoffs and limitations of these capabilities.

The Opportunity: Level the Playing Field with Machine Learning for Cyber Defense

Machine learning promises to augment and automate a range of cybersecurity functions. Areas like automated vulnerability testing—where the security of the ML system itself presents less of a concern—may be most ripe for applying machine learning at scale.[4] This report considers a more speculative question—can ML-enabled defenses engage with attackers in real-time, such as detecting malicious activity or interfering with ongoing attacks, without the need for close human supervision? This is the promise many see in AI and seems to be what NSCAI envisioned when it called for "machine-speed threat detection and mitigation."[5] Before exploring the challenge of robustness, it is worth briefly explaining what machine learning has to offer that justifies this interest.

ML systems find patterns in vast quantities of data that are useful for making predictions in situations of uncertainty. They can reach or exceed human performance in tasks ranging from image classification to complex strategy games. In cybersecurity, this could translate to detection capabilities that significantly raise the bar for attackers. For instance, intrusion detection systems could leverage the vast data on network activity available to defenders to define a baseline of normal behavior, helping defenders spot anomalies more quickly and accurately. Attackers would need to not only avoid obvious red flags, but mirror legitimate activity at a granular level. Malware detection similarly benefits from systems that can churn through large datasets to discover broader patterns that distinguish malicious from benign code. Cybersecurity services already commonly employ machine learning to help analyze malware, such as identifying clusters of similar malicious samples and matching them with known malware. While traditional antivirus systems struggle to keep up as attackers update their code, ML-enabled systems will (ideally) be able to detect unseen malware by identifying deeper patterns that characterize malicious code.[6]

Machine learning holds the potential to automate not only detection, but also defensive responses to attacks. This includes, in theory, measures that can dynamically adapt to interfere with and mitigate attacks. Decoy data or networks known as "honeypots" have long been employed, but machine learning could tailor them to ongoing attacks, more effectively luring attackers and coaxing them into revealing their capabilities.[7] Researchers are experimenting with systems that could automatically reconfigure networks on the fly to impede an attacker's operation.[8]

These defensive applications could correct the long-standing, asymmetric advantages attackers have enjoyed. These include attackers' ability to carefully plan operations and surprise defenders, maneuver through often predictable target environments once inside a network, and continuously adapt capabilities to keep ahead of defenses. ML-enabled defenses might be able to anticipate future attacks and react instantly. They could realize defenders' latent "home field advantage," turning surprise and deception against attackers.[9]

The true test for machine learning, however, is whether it can contend with offensive *campaigns* in which attackers persist and adapt their tactics, including by targeting flaws in the ML model itself. The most sophisticated threat actors will harness machine learning themselves for offensive applications.[10] ML-based antivirus systems may have to contend with malware augmented to better evade detection.[11] Similarly, attackers may trick intrusion detection systems with ML capabilities that learn how to conceal attacks in a target environment or disguise communication with command and control servers.[12] Whether machine learning can level the playing field for defenders thus depends crucially on its ability to withstand systematic efforts to defeat it.

The Problem: Making Machine Learning Defenses Robust

Even the most sophisticated ML systems are often vulnerable to deception. An attacker could evade an ML-based malware classifier by camouflaging malware as a legitimate file or fool an intrusion detection system by mimicking normal user behavior. ML defenses must therefore be *adversarially robust*, meaning they continue to perform reliably in the face of such attempts at deception.* They must also be *accurate* in their performance at a given task. An inaccurate malware classifier might only correctly identify, say, half of incoming malware samples as malicious. An accurate but non-robust malware classifier may spot 99 percent of malware but fail completely in the face of a clever deception. This is a subtle but important distinction, because these goals—accuracy and adversarial robustness—may be fundamentally in conflict. Understanding this tradeoff requires a closer look at what makes machine learning vulnerable.

Vulnerabilities in Machine Learning

Machine learning looks for patterns or statistical regularities in data that are useful for making predictions. To maximize the accuracy of its predictions, a system will look for any patterns that are useful, regardless of whether they could lead to error. Consider, for instance, an image classifier that learns to distinguish wolves from huskies based on whether there is snow in the image.[13] This may be the most efficient way to accurately classify a training dataset filled with many images of wolves with snow and huskies without snow. But this learned association leaves the system vulnerable to error when confronted with a more representative training set or with a deliberate deception. An attacker able to interact with the system could discover this flaw and, using readily accessible techniques, create a deceptive input, for instance, by adding a

* Robustness more broadly refers to the ability of an ML system to perform reliably when conditions deviate from those under which the system was trained, including from changes in the distribution of data in the environment and adversarial manipulation of inputs. See Tim G. J. Rudner and Helen Toner, "Key Concepts in AI Safety: An Overview" (Center for Security and Emerging Technology, March 2021), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-an-overview/.

snow-like pattern to an image of a husky. Research on machine learning is replete with such “adversarial examples”—deceptive inputs that researchers create to change the model’s prediction.[14] A growing body of literature demonstrates this type of attack on a range of cybersecurity applications, from spam filters to malware and intrusion detection systems.[15]

What makes a system vulnerable to these deceptions isn’t simply an error written into the code, as is often the case with software vulnerabilities. Rather, it stems from the fact that ML systems rely on identifying correlations rather than understanding causal relationships, and data is often full of spurious correlations like “snow” and “wolf” that are useful rules of the thumb but not always accurate. Indeed, some researchers describe adversarial examples as “not bugs, but features” of machine learning because they demonstrate that the system has learned a pattern useful for prediction.[16] Even though a manipulated image can trick it, the system is doing exactly what it’s supposed to do—making predictions based on associations.

Because they are, to some degree, inherent to how ML systems function, there’s no easy way to prevent these vulnerabilities from emerging. In practice, it can be extremely difficult to see when a system learns a spurious association, such as “snow” and “wolf.” The patterns that ML systems discover are often imperceptible and counterintuitive to humans—which is precisely the strength of machine learning. And developers can’t simply ask the system—machine learning, specifically deep learning, is generally resistant to querying by developers. This makes it hard to figure out why a deceptive image tricks the system. Nor is it feasible to identify vulnerabilities by testing every possible combination of inputs that might trick a system, because the number can be astronomical.

These challenges motivate the search for reliable ways to defend against adversarial examples—but to little avail. Surveying the state of ML security in 2017, Ian Goodfellow and Nicolas Papernot described defensive innovations as “playing a game of whack-a-mole: they close some vulnerabilities, but leave others open.”[17] Rather than addressing the underlying source of the vulnerability, these defenses appear to merely mitigate one or another specific attack methodology. In a study published in February 2020, a group of researchers defeated 13 different defensive techniques that had been presented to leading conferences on ML security.[18] The group found ways to beat each defense despite the fact that each had been tested against repeated attacks—demonstrating not only the weaknesses of the defenses but also the insufficiency of evaluations conducted on them. Reflecting on their findings, one expert questioned whether defenses against adversarial examples were improving at all.[19]

The Accuracy-Robustness Tradeoff

Researchers have found various ways to weed out spurious associations to produce a model more robust to attack. However, doing so appears to come at the cost of the overall accuracy of the model.[20] This seems to be because those features were useful for making predictions in situations of uncertainty. Tasks like distinguishing wolves from huskies can be arduous for a machine. The developer can craft data specifically to train the system not to rely on snow as an indicator, but without this indicator it's harder for the system to identify wolves and huskies in the first place. In other words, the system might be less susceptible to deception, but it's less effective at its main task.

Sacrificing some accuracy for robustness might be worthwhile. But in some cases, it would present a dilemma. Consider a hypothetical case involving a self-driving car:[21] A developer might have to choose between a system with a risk of crashing once in a million miles (under normal conditions) and one that crashes once every hundred thousand miles but is more adversarially robust. The latter might be more robust because it avoids relying on certain patterns, such as reflections of light in certain weather conditions, making it harder for an attacker to fool but increasing the risk of error under those conditions. The developer might reasonably choose the former system if they assess a low risk that some actor has both means and motive to orchestrate an attack. If the threat of an adversarial attack were significantly higher, it might make more sense to sacrifice some accuracy for the sake of robustness.

But what if adversary behavior is the very thing the model aims to predict? In such cases, the accuracy-robustness tradeoff becomes a choice to prioritize between different kinds of malicious threats. For instance, an ML system trained to detect communication between malware and malicious command and control servers might be vulnerable to malware traffic altered to evade the model. The developer can train the system with adversarial examples to harden it against such attacks, but this might actually make it worse at detecting *unaltered* malware traffic.[22]

Maximizing the accuracy of an antivirus system might improve its overall detection rate while inadvertently making it more susceptible to a deceptive attack, such as an attempt to camouflage malware as a legitimate file. Researchers have already demonstrated a successful attack against a highly accurate, deployed ML-enabled antivirus engine.[23] The researchers reverse-engineered the system, discovering that the model had learned to strongly associate certain sequences of characters with benign files. They simply had to attach the sequence to a malicious file to trick the system into classifying it as benign. A system carefully designed to prevent such blind spots might be robust to this kind of deception, but more susceptible in general to false negatives (malware that slips by the system) or false positives (legitimate files mistakenly flagged as malicious).

What if a commercial antivirus service like the one described above protects an important government network? The vendor might continuously update the ML system with training data from a wide range of clients' networks. If the vendor simply prioritizes accuracy or minimizing false positives, the system may learn a spurious association that creates a way to bypass the defense. In any case, what happens on completely unrelated networks shapes the system in a way that directly impacts the security of the government network.

The Continuous Balancing Act for ML-Cyber Defense

The purpose of drawing attention to this tradeoff is to underscore the complex challenge of machine learning security. Unlike simply patching a software vulnerability, measures to secure an ML system often change how it responds to inputs in ways that are difficult to predict. Balancing tradeoffs between different risks to a system can be difficult even under favorable conditions. In the face of constantly evolving cyber threats, it will be a particularly acute problem for three reasons:

First, **attackers can constantly probe defenses** to search for vulnerabilities in the ML system. With enough attempts, they are likely to succeed in evading even highly accurate systems.[24] This makes the deployment environment uniquely onerous for machine learning, even compared to other adversarial contexts like autonomous weapons systems, where engagements are episodic and capabilities are slower to evolve. Moreover, the cybersecurity arena features many actors observing and learning from one another’s successes and failures. Because a deceptive input that fools one ML model will often fool other models trained for the same task, an attacker may learn useful lessons from repeated attacks on one ML system for crafting a deception to defeat others.[25] In other words, defenders can’t assume that just because an attacker hasn’t probed their specific ML system the attacker can’t find a way to defeat it.

Second, **attackers, by definition, shape the data** used to train systems to detect malicious behavior. An attacker could “poison” a machine learning system by feeding in data that would lead it to learn associations that could expose it to future attack. For instance, an attacker might try and mistrain an intrusion detection system by habituating the system to the attacker’s presence in a network.[26] ML systems continuously learning in deployment must be able to adapt to adversary behavior without maladapting to adversary deceptions.

Third, compounding these factors is the fact that **offensive capabilities and the environment itself constantly evolve**. This isn’t a stable problem like distinguishing wolves from huskies. The data useful for describing normal network behavior or common malware threats quickly becomes outdated. Deploying ML-cyber defenses at scale against persistent threats may require them to learn and adapt while deployed. That means constantly taking in new training data to update models while attackers try to outmaneuver them or actively corrupt the adaptation process.

Research on ML security often treats robustness as a static problem, distinct from the challenges of learning while deployed in a dynamic environment.[27] In cybersecurity, these problems collide, creating tension between the need for continuous learning and adaptation to keep pace with malicious threats in general and measures to ensure robustness against deceptions by adaptive attackers. Balancing these risks will grow increasingly onerous for defenders as they rely on machine learning at scale.

Taking Stock: The Outlook for ML-Cyber Defense

What does this balancing act mean for the viability of ML-cyber defense at scale? We are still learning about the tradeoff between accuracy and robustness in machine learning, and more so about how it might manifest in cybersecurity.[28] Studies of cybersecurity applications comprise only a tiny fraction of research on ML robustness. Still, some view cyber defense as a dead end for machine learning. Experts argue that machine learning simply won't work reliably for dynamic problems like detecting attacks.[29]

While the skepticism toward machine learning is reasonable, there is nevertheless a strategic interest in pursuing secure and reliable machine learning for cyber defense. First, the status quo in cybersecurity, broadly speaking, isn't working. Machine learning can fill gaps where traditional approaches have failed, such as detecting metamorphic malware that evolves to evade signature-based antivirus systems.[30] Second, it may be necessary for the United States to compete against top-tier adversaries who augment their offensive operations with AI. Concerns about robustness are far less likely to inhibit offensive applications of machine learning. Adversaries could use it to enable stealthier and faster attacks, undermining traditional detection techniques.[31]

Moreover, a range of avenues for further research will better equip defenders to manage the tradeoffs of machine learning. One such avenue includes researching ways to take advantage of the constraints under which attackers must operate in order to achieve their objectives.[32] Malware, for instance, needs to be able to carry out malicious functions, and tweaking it to evade detection may at some point alter its functionality. The adage that attackers "only need to be right once" might apply when finding a vulnerability in a malware classifier in a lab setting, but in the context of an actual offensive operation, the attacker has to succeed in a series of steps. Attacks on ML-based malware classifiers demonstrated by researchers in experimental settings rarely constitute a real-world threat because they fail to maintain functionality or simply get caught by other modes of detection.[33] Defenders could combine ML-based static analysis of a file with non-ML techniques such as dynamic analysis of the file executed in a "sandbox"—an isolated environment in which the defender can safely observe a suspicious file's behavior.[34] This is why attacks on ML models generally aren't seen as a significant threat by vendors who deploy them in limited roles in concert with other tools and techniques.[35] An ML system doesn't need to successfully detect all attacks so long as the attacks it fails to detect would be caught by other means.

A second avenue for further research includes techniques that enable defenders to make more deliberate choices about accuracy-robustness tradeoffs. One approach already employed for cybersecurity is to create an "ensemble" of differently trained models, so that if one fails to spot an attack, others might catch it.[36] However, employing multiple models is not sufficient for a strong defense. Differently trained models may discover the same spurious correlations. Defenders need better ways to gauge and certify a model's robustness to specific threats. For instance, developers can design a "monotonic" classifier that only learns the positive indicators of a malicious file rather than features of both malicious and benign files.[37] This makes it impossible for an attacker to use features associated with benign files to camouflage malware, though other methods of evasion might work. This method still produces a less accurate system overall, but it enables defenders to be more cognizant about which threats the model is robust against, so it can be employed more effectively in concert with other defenses. Microsoft's antivirus service already employs three monotonic models together with other types of defensive systems.[38] Finally, researchers are developing methods of adversarial training with less negative impact on accuracy.[39]

Defenders have options to manage this tradeoff at the tactical level. The challenge will be figuring out how to use them proactively and effectively in a threat landscape that will evolve in response to the deployment of ML defenses at scale. Specifically, three problems for defenders will emerge as cybersecurity becomes more reliant on ML systems:

1) Understanding and managing an expanding attack surface: Machine learning will create new dependencies on data sources, open-source tools, and commercial services that introduce new attack vectors cutting across the ML and cybersecurity ecosystem. For instance, an ML-based intrusion detection system might consist of one “global” model receiving updates from numerous models deployed locally on clients’ networks and, in turn, periodically updating those local models (a process referred to as federated learning). These feedback loops create new vulnerabilities. An attacker could hack local models in order to poison the global model, which could then expose other locally deployed models.[40] Defenders in government and critical infrastructure settings, in particular, need to understand the risks and benefits of such dependencies. Major cybersecurity vendors such as CrowdStrike rake in vast troves of data on threats to feed to their ML systems.[41] They may have some of the most effective capabilities, but a vendor dealing primarily with opportunistic threat actors may optimize their systems against a wide spectrum of threats in a way that leaves them vulnerable to a targeted attack by a sophisticated adversary.

2) Anticipating how offensive campaigns will react and evolve: Attackers will not only adapt their tactics to find flaws in a model, they will adapt their strategies to circumvent or sabotage ML-cyber defenses.[42] An expanding attack surface will provide opportunities to poison training data to corrupt a model. An attacker might only need to tamper with 1 percent of training samples to insert a “backdoor” into a model that would cause it to react to a specific input in a way favorable to the attacker once deployed.[43] The most capable nation–state actors may counter ML defenses with their own applications, such as using machine learning to generate malware that can bypass a defense.[44]

3) Tailoring tools and techniques for robustness to evolving cyber threats: Techniques to improve robustness are only effective if they make the system robust to the real-world threats defenders face. Otherwise, they can even be counterproductive. For instance, generating synthetic adversarial inputs to retrain a system might harden it to a different distribution of attacks than it will encounter in deployment, leaving it more vulnerable to real-world threats.[45] Thus, even though high quantity and quality data is essential, defenders can't simply assume "more data" will produce a robust defense. If their tools aren't guided by broader intelligence on evolving threats, they may only produce a false sense of security.

These problems rise above the level of any one developer or defender. They present a collective challenge that must be addressed in policy and strategy. A failure to understand and manage the risks of machine learning risks repeating the mistakes made with the adoption of information and communications technologies in general: decisions to prioritize openness and efficiency rather than security—from the architecture of the internet itself to the web of hardware and software products operating within it—led to widespread reliance on fundamentally insecure technologies. Because of the failure to anticipate and manage their risks, cybersecurity has been an uphill battle ever since. Now there is an opportunity to shape the trajectory of the ML ecosystem toward a more tenable situation for defenders.

How to Make Machine Learning Work for Cyber Defense

Machine learning will be an essential part of the defender's toolkit. But deploying ML systems securely will require defenders to manage a continuous balancing act. The aim of policy and strategy should be to put defenders on solid ground to navigate machine learning's tradeoffs. This section includes four recommendations for government efforts to shape the trajectory of the emerging ML-cybersecurity ecosystem toward a more tenable situation for defenders:

1) Build security into the process of ML design and development for cybersecurity applications. Generally speaking, ML development has been driven by the goal of maximizing accuracy in the most efficient manner. Cybersecurity demands ML systems that can not only make accurate predictions, but do so reliably under constant stress from changing environmental conditions and adversarial interference. These challenges call for the development of a "dynamic-adversarial learning paradigm" tailored to the cybersecurity context.[46] This is a holistic approach to security as a "cyclic and ongoing process" prioritizing robustness, not just accuracy, from design and training to deployment and updating.[47]

Operationalizing this approach requires research at the intersection of ML security and cybersecurity. This begins with understanding the threat by developing realistic threat models for specific cybersecurity applications.[48] Defenders need to be able to assess the relative threat, for instance, from an adversary that is able to directly probe a deployed system versus an adversary that might only be able to gain insights into a model indirectly. A second area for research, then, is to identify key robustness properties and develop ways to measure and certify them. Specifically, developers need techniques to certify "global" robustness properties that aren't invalidated by changes in a model learning in deployment.[49] One expert compares the current state of machine learning security to cryptography in the 1920s—not only are the most secure systems easily breakable, but researchers lack even the metrics to properly evaluate security.[50] A final area of focus should be the development of broader system-level defenses that detect or prevent attacks that would otherwise defeat the model. These include measures to detect attempts to probe deployed systems and prevent them from "leaking" information to attackers who might try to reverse engineer them.[51] A holistic approach to defense requires both mitigating vulnerabilities in a model and implementing measures to prevent attackers from finding and exploiting those vulnerabilities that remain.

2) Promote resilience through system diversity and redundancy. Cyber defense would benefit from further research on innovative approaches to machine learning that incorporate diversity and redundancy in the design and implementation of systems.

Cybersecurity vendor F-Secure's "Project Blackfin," for instance, seeks to develop multiple ML agents that model different aspects of a network environment and work collaboratively to identify intrusions.[52] By combining multiple models relying on different modalities—or different ways of perceiving—the resulting system may be more robust to deception than an ensemble of models looking at the same data.[53]

Even with better tools to improve robustness, ML systems won't be foolproof. Effective implementation depends on limiting the impacts of a system's failure. As discussed above, vendors already commonly rely on multiple tools and techniques, both ML-based and otherwise. But as reliance on ML systems grows, decisionmakers must establish different thresholds of risk tolerance that guide where and how to rely on ML systems and when to supplement them with non-ML tools and safeguards.

3) Manage risks cutting across the machine learning and cybersecurity ecosystem. Machine learning and cybersecurity are both, generally, collective endeavors that depend on trusted relationships and connections. This can be an advantage for defenders: a common repository such as VirusTotal provides a dataset of malware collectively curated by multiple antivirus vendors. But it could become a liability: an attacker able to inject manipulated data could exploit this dependency to mistrain ML-based antivirus systems. There has already been a reported case of a malicious actor uploading mutated variants of ransomware in what may have been an attempt to poison VirusTotal data.[54] Open-source tools and models relied on by the ML research community present similar opportunities for attack.[55] And the threat from poisoning extends to the process of machine learning in deployment, as in the example of federated learning described above.

Decision makers must ensure that as ML systems are adopted, defenders are mapping and managing the risks from these critical dependencies. In some cases, there may be technical measures that can address these threats.[56] But in others, the consequences of failure might be severe enough that defenders must err on the side of avoiding creating dependencies in the first place.

4) Counter strategic rivals' attempts to compromise and sabotage ML development. Defenders' reliance on machine learning will create strong incentives for adversaries to target the ML development process itself—particularly for nation–state actors seeking to enable future cyber operations. It is far easier to attack an ML system with even partial knowledge of a target model's parameters, architecture, or training data and methods.[57] Adversaries will look for opportunities to gain intelligence on the inner workings of ML systems by acquiring training datasets, infiltrating commercial or open-source projects, or simply by purchasing products to reverse-engineer them.[*] They may even sabotage ML systems by inserting backdoors into models that make their way into deployed defenses.

The success of defense at a tactical level will depend on coordination across government and the private sector to anticipate and thwart offensive campaigns aiming to fatally compromise defenses before they are even deployed. Vendors must carefully vet and secure data and components crucial to the integrity of their services. Government agencies should explore how to extend ongoing efforts to secure supply chains and prevent adversaries from acquiring sensitive technologies and data to include ML capabilities.

Conclusion

AI is no panacea for cybersecurity, but it could become indispensable. Existing ML methods simply weren't designed for security and still less for an environment characterized by constant change and deception. The growing role of machine learning in cybersecurity raises a serious question: how will choices made now about the design and implementation of machine learning shape

[*] The Chinese PLA, for instance, already purchases foreign antivirus systems, likely in order to reverse engineer and find ways to defeat them. Insikt Group, "China's PLA Unit 61419 Purchasing Foreign Antivirus Products, Likely for Exploitation," *Recorded Future*, May 5, 2021, https://www.recordedfuture.com/china-pla-unit-purchasing-antivirus-exploitation/.

the trajectory toward a more or less tenable situation for defenders?

This challenge calls for collaboration that bridges the gaps between the various research and practitioner communities working on different aspects of ML security and cybersecurity. Policymakers, meanwhile, need to consider how to steer the various stakeholders developing and adopting ML capabilities toward prioritizing security, not just efficiency. Proactively managing the problems of ML-cyber defense at scale will lead to a far better situation for defenders in the future.

Authors

Wyatt Hoffman is a research fellow at CSET, where he works on the CyberAI Project.

Acknowledgments

The author would like to thank Christopher Whyte, Sadia Afroz, John Bansemer, and Andrew Lohn for their invaluable feedback and comments on earlier versions of this manuscript.

Document Identifier: doi: 10.51593/2021CA007

Endnotes

[1] National Security Commission on Artificial Intelligence, *Final Report* (Washington, DC: NSCAI, March 2021), https://www.nscai.gov/2021-final-report/.

[2] For instance, former head of Air Combat Command General (ret.) Mike Holmes expressed reservations about relying on intelligence from ML capabilities. Colin Clark, "Air Combat Commander Doesn't Trust Project Maven's Artificial Intelligence — Yet," *Breaking Defense*, August 21, 2019, https://breakingdefense.com/2019/08/air-combat-commander-doesnt-trust-project-mavens-artificial-intelligence-yet/. The NSCAI final report also discusses machine learning robustness extensively.

[3] Dimitris Tsipras et al., "Robustness May Be at Odds with Accuracy," arXiv preprint arXiv:1805.12152 (2018), http://arxiv.org/abs/1805.12152.

[4] For an extensive overview of the current state of play and near-term prospects of machine learning in cybersecurity, see Micah Musser and Ashton Garriott, "Machine Learning and Cybersecurity: Hype and Reality" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/publication/machine-learning-and-cybersecurity/.

[5] National Security Commission on Artificial Intelligence, *Final Report*.

[6] 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; Sean Park, Iqbal Gondal, Joarder Kamruzzaman, and Jon Oliver, "Generative Malware Outbreak Detection" (Trend Micro, 2019), https://documents.trendmicro.com/assets/white_papers/GenerativeMalwareOutbreakDetection.pdf.

[7] 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.

[8] See, for instance, Taha Eghtesad, Yevgeniy Vorobeychik, and Aron Laszka, "Deep Reinforcement Learning Based Adaptive Moving Target Defense," arXiv preprint arXiv:1911.11972 (2019), http://arxiv.org/abs/1911.11972; Andres Molina-Markham, Ransom K. Winder, and Ahmad Ridley, "Network Defense is Not a Game," arXiv preprint arXiv:2104.10262 (2021), https://arxiv.org/pdf/2104.10262.pdf.

[9] On the opportunities for defenders to leverage their "home field advantage" see 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/.

[10] For an overview of the potential offensive applications of machine learning, see Ben Buchanan, John Bansemer, Dakota Cary, Jack Lucas, and Micah Musser, "Automating Cyber Attacks: Hype and Reality" (Center for Security and Emerging Technology, November 2020), https://cset.georgetown.edu/publication/automating-cyber-attacks/.

[11] Researchers have demonstrated proofs-of-concept for the use of ML capabilities to generate attacks on ML systems in cybersecurity. See, for instance, Wei Song et al., "MAB-Malware: A Reinforcement Learning Framework for Attacking Static Malware Classifiers," arXiv preprint arXiv:2003.03100 (2021), https://arxiv.org/pdf/2003.03100.pdf; Luca Demetrio et al., "Functionality-preserving Black-box Optimization of Adversarial Windows Malware," arXiv preprint arXiv:2003.13526 (2021), http://arxiv.org/abs/2003.13526.

[12] Dakota Cary and Daniel Cebul, "Destructive Cyber Operations and Machine Learning" (Center for Security and Emerging Technology, November 2020), https://cset.georgetown.edu/publication/destructive-cyber-operations-and-machine-learning/; Jie Li et al., "Dynamic Traffic Feature Camouflaging via Generative Adversarial Networks," in *2019 IEEE Conference on Communications and Network Security (CNS)* (Washington, DC: IEEE, 2019): 268–276, https://doi.org/10.1109/CNS.2019.8802772.

[13] This example is borrowed from Christoph Molnar, "Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable," *Github*, 2018, https://christophm.github.io/interpretable-ml-book/.

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv preprint arXiv:1412.6572 (2015), http://arxiv.org/abs/1412.6572.

[15] For a recent survey of the literature on ML vulnerabilities in cybersecurity applications, see Ihai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach, "Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain," arXiv preprint arXiv:2007.02407 (2021), http://arxiv.org/abs/2007.02407.

[16] Andrew Ilyas et al., "Adversarial Examples Are Not Bugs, They Are Features," arXiv preprint arXiv:1905.02175 (2019), https://arxiv.org/abs/1905.02175.

[17] 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.

[18] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry, "On Adaptive Attacks to Adversarial Example Defenses," arXiv preprint arXiv:2002.08347 (2020), http://arxiv.org/abs/2002.08347.

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

[20] Tsipras et al., "Robustness May Be at Odds with Accuracy." This tradeoff has been observed in a range of application settings. For example, robotic systems trained with adversarial examples become more robust to attack but also more prone to safety accidents. See Mathias Lechner et al., "Adversarial Training Is Not Ready for Robot Learning," arXiv preprint arXiv:2103.08187 (2021), http://arxiv.org/abs/2103.08187.

[21] This hypothetical example is borrowed from Nicholas Carlini, "Trustworthy AI: Adversarially (non-)Robust ML," in AI For Good 2021 (Geneva: March 25, 2021), *YouTube*, March 31, 2021, https://www.youtube.com/watch?v=qgsmd2LaZA4.

[22] Novo and Morla, for instance, found in a set of experiments that adversarial training improved detection of altered malware traffic but decreased accuracy at detecting unaltered malware traffic. Carlos Novo and Ricardo Morla, "Flow-based Detection and Proxy-based Evasion of Encrypted Malware C2 Traffic," in *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security (AISec'20)* (New York, NY: Association for Computing Machinery, 2020): 83–91. https://doi.org/10.1145/3411508.3421379.

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

[24] 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).

[25] Adversarial examples created to fool one ML model often transfer to other models. However, the transferability of attacks on cybersecurity applications in practice is an open question. Demetrio et al. found that malware generated to evade ML classifiers successfully evaded multiple commercial antivirus systems, while Song et al. found commercial antivirus systems to be more resistant to transfer attacks. Even if attacks don't directly transfer, attackers will likely learn

lessons from attempts against ML systems that carry over to future operations. See Florian Tramèr et al., “The Space of Transferable Adversarial Examples,” arXiv preprint arXiv:1704.03453 (2017), http://arxiv.org/abs/1704.03453; Demetrio et al., “Functionality-preserving Black-box Optimization of Adversarial Windows Malware”; Song et al., “MAB-Malware: A Reinforcement Learning Framework for Attacking Static Malware Classifiers.”

[26] Thien Duc Nguyen, Phillip Rieger, Markus Miettinen, and Ahmad-Reza Sadeghi, “Poisoning Attacks on Federated Learning-based IoT Intrusion Detection System,” in *Proceedings of DISS 2020 – Workshop on Decentralized IoT Systems and Security* (San Diego, CA: Internet Society, 2020), https://www.ndss-symposium.org/wp-content/uploads/2020/04/diss2020-23003-paper.pdf.

[27] Tegjyot Singh Sethi and Mehmed Kantardzic, “When Good Machine Learning Leads to Bad Security: Big Data (Ubiquity Symposium),” *Ubiquity* 18 (May 2018): 1–14.

[28] Among the open questions are whether and how the accuracy-robustness tradeoff might apply to reinforcement learning, a subset of machine learning that could enable more sophisticated automated cyber defenses. RL systems are vulnerable to deceptive attacks like those affecting other ML systems, so RL-based defenses might present similar tradeoffs between general performance and robustness to different types of threats. See Sandy Huang et al., “Adversarial Attacks on Neural Network Policies,” arXiv preprint arXiv:1702.02284 (2017), https://arxiv.org/abs/1702.02284; Yi Han et al., "Reinforcement Learning for Autonomous Defence in Software-Defined Networking,” arXiv preprint arXiv:1808.05770 (2018), http://arxiv.org/abs/1808.05770.

[29] For instance, Thomas Dullien argues that malware and intrusion detection, and other engagements between attacker and defender, present dynamic problems for the defender and are therefore “sketchy” areas for applying machine learning. He suggests instead focusing on other cybersecurity opportunities for machine learning that present more stable problems, such as generating large amounts of decoy data to trick attackers. See Thomas Dullien, “Machine Learning, Offense, and the Future of Automation,” ZeroNights 2017 (Moscow: November 16, 2017), December 26, 2017, *YouTube*, https://www.youtube.com/watch?v=BWFdxAG_TGk.

[30] Musser and Garriott, “Machine Learning and Cybersecurity: Hype and Reality.”

[31] Buchanan et al., “Automating Cyber Attacks: Hype and Reality.”

[32] Ihai Rosenberg et al., “Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain,” arXiv preprint arXiv:2007.02407 (2020), http://arxiv.org/abs/2007.02407.

[33] Sadia Afroz, “Lessons from Adversarially Attacking Commercial Malware Detectors,” in IEEE Symposium on Security and Privacy, “[DLS2021] 4th Deep Learning And Security Workshop – Part 1” (4th Deep Learning and Security Workshop, May 27, 2021), *YouTube*, June 9, 2021, https://www.youtube.com/watch?v=fMN5EjIL9P0.

[34] Musser and Garriott, “Machine Learning and Cybersecurity: Hype and Reality.”

[35] Afroz, “Lessons from Adversarially Attacking Commercial Malware Detectors.”

[36] See, for example, Randy Treit, Holly Stewart, and Jugal Parikh, “Protecting the Protector: Hardening Machine Learning Defenses against Adversarial Attacks,” *Microsoft*, August 9, 2018, https://www.microsoft.com/security/blog/2018/08/09/protecting-the-protector-hardening-machine-learning-defenses-against-adversarial-attacks/.

[37] Inigo Incer, Michael Theodorides, Sadia Afroz, and David Wagner, “Adversarially Robust Malware Detection Using Monotonic Classification,” in *IWSPA’18: 4th ACM International Workshop on Security And Privacy Analytics* (March 19–21, 2018), https://doi.org/10.1145/3180445.3180449.

[38] Unsurprisingly, Microsoft researchers have found that these monotonic models are robust to attacks that try to mask malware using features of legitimate files, but their overall detection rate is significantly lower than a model trained without this constraint. See Geoff McDonald, “New Machine Learning Model Sifts through the Good to Unearth the Bad in Evasive Malware,” *Microsoft*, July 25, 2019, https://www.microsoft.com/security/blog/2019/07/25/new-machine-learning-model-sifts-through-the-good-to-unearth-the-bad-in-evasive-malware/.

[39] See, for instance, Qi-Zhi Cai, Chang Liu, and Dawn Song, “Curriculum Adversarial Training,” *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (IJCAI-18) (2018), https://www.ijcai.org/proceedings/2018/0520.pdf.

[40] Alexey Kirichenko, David Karpuk, and Samuel Marchal, “How to attack distributed machine learning via online training,” F-Secure, October 6, 2020, https://labs.f-secure.com/blog/how-to-attack-distributed-machine-learning-via-online-training/.

[41] National Academies of Sciences, Engineering, and Medicine, *Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop*.

[42] A previous CSET report explores how the deployment of ML-cyber defenses might change the dynamics of cyber operations. See Wyatt Hoffman, "AI and the Future of Cyber Competition" (Center for Security and Emerging Technology, January 2021), https://cset.georgetown.edu/publication/ai-and-the-future-of-cyber-competition/.

[43] Giorgio Severi et al., "Exploring Backdoor Poisoning Attacks Against Malware Classifiers," arXiv preprint arXiv:2003.01031 (2020), http://arxiv.org/abs/2003.01031.

[44] Several studies have demonstrated such capabilities, such as Song et al., "MAB-Malware: A Reinforcement Learning Framework for Attacking Static Malware Classifiers."

[45] Techniques like adversarial training harden a system to specific types of attacks. Researchers have found that in some cases, hardening a defense against one type of deception actually makes it more vulnerable to other types of deceptions. Florian Tramèr and Dan Boneh, "Adversarial Training and Robustness for Multiple Perturbations," arXiv preprint arXiv:1904.13000 (2019), https://arxiv.org/abs/1904.13000.

[46] See Tegjyot Singh Sethi and Mehmed Kantardzic, "When Good Machine Learning Leads to Bad Security: Big Data (Ubiquity Symposium)," *Ubiquity* 18 (May 2018): 1–14.

[47] Sethi and Kantardzic, "When Good Machine Learning Leads to Bad Security: Big Data (Ubiquity Symposium)."

[48] There are important ongoing efforts to model threats to ML systems in general—notably MITRE's Adversarial Threat Landscape for Artificial-Intelligence Systems. This provides a foundation to model threats specific to cybersecurity applications. See "ATLAS," The MITRE Corporation, available at https://atlas.mitre.org/.

[49] Yizheng Chen et al., "Learning Security Classifiers with Verified Global Robustness Properties," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)* (November 15–19, 2021), https://arxiv.org/pdf/2105.11363.pdf.

[50] 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.

[51] See, for instance, Steven Chen, Nicholas Carlini, and David Wagner, "Stateful Detection of Black-Box Adversarial Attacks," in *SPAI '20: Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence* (October 2020), https://dl.acm.org/doi/10.1145/3385003.3410925; Tegjyot Singh Sethi et al., "A Dynamic-Adversarial Mining Approach to the Security of Machine Learning," arXiv preprint arXiv:1803.09162 (2018), http://arxiv.org/abs/1803.09162.

[52] "Project Blackfin," F-Secure, 2021, https://www.f-secure.com/en/about-us/research/project-blackfin.

[53] For further discussion of the potential benefits of employing diverse systems relying on different modalities, see Javier Ideami, "Towards the End of Deep Learning and the Beginning of AGI," *Towards Data Science*, March 17, 2021, https://towardsdatascience.com/towards-the-end-of-deep-learning-and-the-beginning-of-agi-d214d222c4cb.

[54] See the VirusTotal poisoning case study in "ATLAS," The MITRE Corporation.

[55] Andrew Lohn, "Poison in the Well: Securing the Shared Resources of Machine Learning" (Center for Security and Emerging Technology, June 2021), https://cset.georgetown.edu/publication/poison-in-the-well/.

[56] Nguyen et al., for instance, propose a defense against poisoning threats to federated learning-based intrusion detection systems. Thien Duc Nguyen et al., "FLGUARD: Secure and Private Federated Learning," arXiv preprint arXiv:2101.02281 (2021), http://arxiv.org/abs/2101.02281.

[57] 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.