ARTIFICAL INTELLIGENCE AND CYBERSECURITY:
RISING CHALLENGES AND PROMISING DIRECTIONS

WRITTEN STATEMENT OF ERIC HORVITZ
CHIEF SCIENTIFIC OFFICER, MICROSOFT CORPORATION

BEFORE THE U.S. SENATE ARMED SERVICES
SUBCOMMITTEE ON CYBERSECURITY

HEARING ON ARTIFICIAL INTELLIGENCE APPLICATIONS TO OPERATIONS IN CYBERSPACE

MAY 3, 2022

Chairman Manchin, Ranking Member Rounds, and Members of the Subcommittee, thank you for the opportunity to share insights about the impact of artificial intelligence (AI) on cybersecurity. I applaud the Subcommittee for its foresight and leadership in holding a hearing on this critically important topic. Microsoft is committed to working collaboratively with you to help ensure new advances in AI and cybersecurity benefit our country and society more broadly.

My perspective is grounded in my experiences working across industry, academia, scientific agencies, and government. As Microsoft's Chief Scientific Officer, I provide leadership and perspectives on scientific advances and trends at the frontiers of our understandings, and on issues and opportunities rising at the intersection of technology, people, and society. I have been pursuing and managing research on principles and applications of AI technologies for several decades, starting with my doctoral work at Stanford University. I served as a Commissioner on the National Security Commission on AI (NSCAI), was president of the Association for the Advancement of Artificial Intelligence (AAAI), chaired the Section on Computing, Information, and Communication of the American Association for the Advancement of Science (AAAS). I am a member of the National Academy of Engineering (NAE) and the American Academy of Arts and Sciences. I currently serve on the President's Council of Advisors on Science and Technology (PCAST) and on the Computer Science and Telecommunications Board (CSTB) of the National Academies of Sciences.

I will cover in my testimony four key areas of attention at the intersection of AI and cybersecurity that warrant deeper understanding and thoughtful action:

- Advancing cybersecurity with AI
- Uses of AI to power cyberattacks
- Vulnerabilities of AI systems to attacks
- Uses of AI in malign information operations

Before covering these topics, I will provide brief updates on the cybersecurity landscape and on recent progress in AI. I'll conclude my testimony with reflections about directions.

1. Cybersecurity's changing landscape

Attacks on computing systems and infrastructure continue to grow in complexity, speed, frequency, and scale. We have seen new attack techniques and the exploitation of new attack surfaces aimed at disrupting critical infrastructure and accessing confidential data.[1] In 2021 alone, the Microsoft 365 Defender suite, supported by AI techniques, blocked more than 9.6 billion malware threats, 35.7 billion phishing and malicious emails, and 25.6 billion attempts to hijack customer accounts targeting both enterprise and consumer devices.[2,3] Multiple independent reports have characterized the nature and status of different forms of cyberattack.[4] As detailed in Microsoft's recent Digital Defense Report,[5] cyber criminals and nation-state actors continue to adapt their techniques to exploit new vulnerabilities and counter cyber defenses.

[1] https://www.microsoft.com/security/blog/2021/12/15/the-final-report-on-nobeliums-unprecedented-nation-state-attack/

[2] https://news.microsoft.com/wp-content/uploads/prod/sites/626/2022/02/Cyber-Signals-E-1-218.pdf, page 3

[3] https://www.microsoft.com/en-us/research/group/m365-defender-research/

[4] 2018-Webroot-Threat-Report_US-ONLINE.pdf

[5] Microsoft Digital Defense Report, October 2021

To help mitigate these concerning trends, the U.S. government has taken significant steps forward to secure our cyber ecosystem. Congress enacted several recommendations that came out of the Cyberspace Solarium Commission, such as creating the Office of the National Cyber Director and enacting cyber incident reporting legislation. Almost a year ago, the Administration issued Executive Order (E.O.) 14028, *Improving the Nation's Cybersecurity*, which directs agencies to develop and implement a variety of initiatives to raise the bar on cybersecurity across areas, such as supply chain security, and requiring agencies to adopt a zero-trust model. Microsoft has worked diligently to meet deadlines specified in the E.O. on cybersecurity and we support these efforts to encourage a cohesive response to evolving cyber threats.

We expect to face continuing efforts by creative and tireless state and non-state actors who will attempt to attack computing systems with the latest available technologies. We need to continue to work proactively and reactively to address threats and to note changes in systems, technologies, and patterns of usage. On the latter, cybersecurity challenges have been exacerbated by the increasing fluidity between online work and personal activities as daily routines have become more intertwined.[6] The large-scale shift to a paradigm of hybrid work coming with the COVID-19 pandemic has moved workers further away from traditional, controlled environments. Cybersecurity solutions must enable people to work productively and securely across various devices from a variety of non-traditional locations.

2. Advancements in Artificial Intelligence

Artificial intelligence is an area of computer science focused *on developing principles and mechanisms to solve tasks that are typically associated with human cognition*, such as perception, reasoning, language, and learning. Numerous milestones have been achieved in AI theory and applications over the 67 years since the phrase “artificial intelligence” was first used in a funding proposal that laid out a surprisingly modern vision for the field.[7]

Particularly stunning progress has been made over the last decade, spanning advances in machine vision (e.g., object recognition), natural language understanding, speech recognition, automated diagnosis, reasoning, robotics, and machine learning—procedures for learning from data. Many impressive gains across subdisciplines of AI are attributed to a machine learning methodology named *deep neural networks* (DNNs). DNNs have delivered unprecedented accuracy when fueled by large amounts of data and computational resources.

Breakthroughs in accuracy include performances that exceed human baselines for a number of specific benchmarks, including sets of skills across vision and language subtasks. While AI scientists remain mystified by the powers of human intellect, the rate of progress has surprised even seasoned experts.

Jumps in core AI capabilities have led to impressive demonstrations and real-world applications, including systems designed to advise decision makers, generate textual and visual content, and to provide new forms of automation, such as the control of autonomous and semi-autonomous vehicles.

[6] https://www.microsoft.com/security/blog/2021/05/12/securing-a-new-world-of-hybrid-work-what-to-know-and-what-to-do/

[7] J. McCarthy, J., M.L. Minsky, N. Rochester, N., C.E. Shannon, C.E. A Proposal for the Dartmouth Summer Project on Artificial Intelligence, Dartmouth University, May 1955. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

AI technologies can be harnessed to inject new efficiencies and efficacies into existing workflows and processes. The methods also can be used to introduce fundamentally new approaches to standing challenges. When deployed in a responsible and insightful manner, AI technologies can enhance the quality of the lives of our citizenry and add to the vibrancy of our nation and world. For example, AI technologies show great promise in enhancing healthcare via providing physicians with assistance on diagnostic challenges, guidance on optimizing therapies, and inferences about the structure and interaction of proteins that lead to new medications.

AI advances have important implications for the Department of Defense, our intelligence community, and our national security more broadly. Like any technology, the rising capabilities of AI are available to friends and foes alike. Thus, in addition to harnessing AI for making valuable contributions to people and society, we must continue to work to understand and address the possibilities that the technologies can be used by malevolent actors and adversaries to disrupt, interfere, and destroy. AI has important implications for cybersecurity as the technologies can provide both new powers for defending against cyberattacks and new capabilities to adversaries.

3. Advancing Cybersecurity with AI

The value of harnessing AI in cybersecurity applications is becoming increasingly clear. Amongst many capabilities, AI technologies can provide automated interpretation of signals generated during attacks, effective threat incident prioritization, and adaptive responses to address the speed and scale of adversarial actions. The methods show great promise for swiftly analyzing and correlating patterns across billions of data points to track down a wide variety of cyber threats of the order of seconds. Additionally, AI can continually learn and adapt to new attack patterns—drawing insights from past observations to detect similar attacks that occur in the future.

3.1 Assisting and Complementing Workforce

The power of automation and large-scale detection, prioritization, and response made possible by AI technologies can not only relieve the burden on cybersecurity professionals but also help with the growing workforce gap. On the challenges to current cyber workforce: the U.S. Bureau of Labor Statistics estimates cybersecurity job opportunities will grow 33% from 2020 to 2030—more than six times the national average.[8]However, the number of people entering the field is not keeping pace. There is a global shortage of 2.72 million cybersecurity professionals, according to the 2021 (ISC)^2 Cybersecurity Workforce Study released in October 2021.[9]

Organizations that prioritize cybersecurity run security operations teams 24/7. Still, there are often far more alerts to analyze than there are analysts to triage them, resulting in missed alerts that evolve into breaches. Trend Micro released a survey in May 2021 of security operations center decision makers that showed that 51% feel their team is overwhelmed with the overall volume of alerts, 55% are not confident in their ability to efficiently prioritize and respond to alerts, and that 27% of their time is spent dealing with false positives.[10]

[8] https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm

[9] https://www.isc2.org/News-and-Events/Press-Room/Posts/2021/10/26/ISC2-Cybersecurity-Workforce-Study-Sheds-New-Light-on-Global-Talent-Demand

[10] https://newsroom.trendmicro.com/2021-05-25-70-Of-SOC-Teams-Emotionally-Overwhelmed-By-Security-Alert-Volume

AI technologies enable defenders to effectively scale their protection capabilities, orchestrate and automate time-consuming, repetitive, and complicated response actions. These methods can enable cybersecurity teams to handle large volumes of classical threats in more relevant time frames with less human intervention and better results. Such support with scaling on the essentials can free cybersecurity professionals to focus and prioritize on those attacks that require specialized expertise, critical thinking, and creative problem solving. However, additional attention should also be given to general cybersecurity training, security awareness, secure development lifecycle practices, and simulated training modules, including using AI to run intelligent and personalized simulations.

3.2 AI at Multiple Stages of Security

Today, AI methods are being harnessed across all stages of security including *prevention, detection, investigation and remediation, discovery and classification, threat intelligence,* and *security training and simulations*. I will discuss each of these applications in turn.

Prevention. Prevention encompasses efforts to reduce the vulnerability of software to attack, including user identities and data, computing system endpoints, and cloud applications. AI methods are currently used in commercially available technologies to detect and block both known and *previously unknown* threats before they can cause harm. In 2021, AV-Test Institute observed over 125 million *new* malware threats.[11] The ability of machine learning techniques to generalize from past patterns to catch new malware variants is key to being able to protect users at scale.

As an example, last year Microsoft 365 Defender successfully blocked a file that would later be confirmed as a variant of the GoldMax malware. Defender *had never seen the new variant* of GoldMax. The malware was caught and blocked leveraging the power of an AI pattern recognizer working together with a technology known as "fuzzy hashing"—a means for taking a fingerprint of malware.[12] It is important to note that GoldMax is malware that *persists* on networks, feigning to be a "scheduled task" by impersonating the activities of systems management software. Such hiding out as a scheduled task is part of the tools, tactics, and procedures of NOBELIUM, the Russian state actor behind the attacks against SolarWinds in December 2020 and which the U.S. government and others have identified as being part of Russia's foreign intelligence service known as the SVR.

In other work, we have found that AI methods can improve our ability to detect sophisticated *phishing* attacks. Phishing attacks center on *social engineering*, where an attacker creates a fake webpage or sends a fraudulent message designed to trick a person into revealing sensitive data to the attacker or to deploy malicious software on the victim's device, such as ransomware. To help protect people from harmful URLs, AI pattern recognizers have been deployed in browsers and other applications as part of their security services. AI methods can improve detection while lowering false positive rates, which can frustrate end users.[13]

[11] https://www.av-test.org/en/statistics/malware/

[12] https://www.microsoft.com/security/blog/2021/07/27/combing-through-the-fuzz-using-fuzzy-hashing-and-deep-learning-to-counter-malware-detection-evasion-techniques

[13] https://www.microsoft.com/en-us/research/publication/urltran-improving-phishing-url-detection-using-transformers/

Detection. Detection involves identifying and alerting suspicious behaviors *as they happen*. The goal is to *quickly respond to attacks*, including identifying the scale and scope of an attack, closing the attacker's entry, and remediating footholds that the attacker may have established. The key challenge with detecting suspicious activity is to find the right balance between providing enough coverage via seeking high rates of accurate security alerts versus false alarms. AI methods are being leveraged in detection to (1) *triage* attention to alerts about potential attacks, (2) identify multiple attempts at breaches over time that are part of larger and lengthier *attack campaigns,* (3) *detecting fingerprints of the activities of malware* as it operates within a computer or on a network*, (4) identifying the flow of malware* through an organization,[14] and (5) guiding *automated approaches to mitigation* when a response needs to be fast to stop an attack from propagating. For example, an automated system can shut down network connectivity and contain a device if a sequence of alerts is detected that is known to be associated with ransomware activity like the way a bank might decline a credit card transaction that appears fraudulent.

There are several technologies available today to help detect attacks. I will use Microsoft 365 Defender capabilities as an example. A set of neural network models are used to detect a potential attack underway *by fusing multiple signals* about *activities* within a computing system, including processes being started and stopped, files being changed and renamed, and suspicious network communication.[15, 16] In addition, probabilistic algorithms are used to detect high likelihoods of "lateral movement" on a network.[17] Lateral movement refers to malware, such as ransomware, moving from machine to machine as it infects an organization. The goal is to detect signals of concerning patterns of spread and to shut down the infection by isolating potentially infected machines and alerting security experts to investigate. As numerous legitimate operations can appear like lateral movement of malware, simplistic approaches can have high false-positive rates. AI systems can help to raise the rate of capture and block these spreading infections, while reducing false positives.[18]

As a recent example, in March 2022, Microsoft leveraged its AI models to identify an attack attributed to a Russian actor that Microsoft tracks as *Iridium*, also referred to as *Sandworm.* The US government has attributed Iridium activity to a group allegedly based at GRU Unit 74455 of the Main Directorate of the General Staff of the Armed Forces of the Russian Federation. The actor deployed *wiper* malware at a Ukrainian shipping company based in Lviv. Wiper malware erases data and programs on the computers that it infects. The first documented encounter of this malware was on a system running Microsoft Defender with Cloud Protection enabled. The ensemble of machine learning models in Defender, combined with signals across client and cloud, allowed Microsoft to block this malware at first sight.

Investigation and remediation. Investigation and remediation are methods used following a breach to provide customers with a holistic understanding of the security incident, including the extent of the breach, which devices and data were impacted, how the attack propagated through the customer environment, and to

[14] https://dl.acm.org/doi/10.1145/3471621.3471858

[15] https://www.microsoft.com/security/blog/2020/07/23/seeing-the-big-picture-deep-learning-based-fusion-of-behavior-signals-for-threat-detection/

[16] https://www.microsoft.com/security/blog/2020/08/27/stopping-active-directory-attacks-and-other-post-exploitation-behavior-with-amsi-and-machine-learning/

[17] https://www.microsoft.com/security/blog/2019/12/18/data-science-for-cybersecurity-a-probabilistic-time-series-model-for-detecting-rdp-inbound-brute-force-attacks/

[18] https://www.microsoft.com/security/blog/2020/06/10/the-science-behind-microsoft-threat-protection-attack-modeling-for-finding-and-stopping-evasive-ransomware/

seek attribution for the threat.[19] Gathering and doing synthesis from telemetry sources is tedious. Efforts to date include multiple tools to collect telemetry from within and across organizations. The use of AI for investigation and remediation is a promising and open area of research.[20,21]

Threat intelligence. Threat intelligence enables security researchers to stay on top of the current threat landscape by tracking active malicious actors, at times deliberately engaging with them and studying their behavior. Today, Microsoft actively tracks 40+ active nation-state actors and 140+ threat groups across 20 countries.[22,23] AI methods help to identify and tag entities from multiple feeds and intelligence sharing across agencies. AI models show promise with their ability to learn and make inferences about high-level relationships and interactions by identifying similarities across different campaigns for enhancing threat attribution.[24,25]

Recommendations: Advance development and application of AI methods to defend against cyberattacks

- Follow best practices in cybersecurity hygiene, including implementation of core protections such as multifactor authentication. Bolster security teams, regularly test backups and update patches, test incident response plans, and limit internet access to networks that do not require internet connectivity.
- Invest in training and education to strengthen the U.S. workforce in cybersecurity, including education and training programs on cybersecurity for both traditional and AI systems.
- Invest in R&D on harnessing machine learning, reasoning, and automation to detect, respond, and protect every step of the cyberattack kill chain.
- Incentivize the creation of cross-sector partnerships to catalyze sharing and collaboration around cybersecurity experiences, datasets, best practices, and research.
- Develop cybersecurity-specific benchmarks and leaderboards specific to validate research and accelerate learnings.

4. AI-powered cyberattacks

While AI is improving our ability to detect cybersecurity threats, organizations and consumers will face new challenges as cybersecurity attacks increase in sophistication. To date, adversaries have commonly employed software tools *in a manual manner* to reach their objectives. They have been successful in exfiltrating sensitive data about American citizens, interfering with elections, and distributing propaganda on social media

[19] https://www.microsoft.com/security/blog/2021/12/02/structured-threat-hunting-one-way-microsoft-threat-experts-prioritizes-customer-defense/

[20] https://www.microsoft.com/security/blog/2020/07/09/inside-microsoft-threat-protection-correlating-and-consolidating-attacks-into-incidents/

[21] https://www.microsoft.com/security/blog/2020/07/29/inside-microsoft-threat-protection-solving-cross-domain-security-incidents-through-the-power-of-correlation-analytics/

[22] https://www.microsoft.com/security/blog/2022/02/03/cyber-signals-defending-against-cyber-threats-with-the-latest-research-insights-and-trends/

[23] https://www.microsoft.com/security/blog/2021/05/12/securing-a-new-world-of-hybrid-work-what-to-know-and-what-to-do/

[24] https://www.microsoft.com/security/blog/2021/04/01/automating-threat-actor-tracking-understanding-attacker-behavior-for-intelligence-and-contextual-alerting/

[25] https://dl.acm.org/doi/pdf/10.1145/3448016.3452745

without the sophisticated use of AI technologies. [26,27,28] While there is scarce information to date on the active use of AI in cyberattacks, it is widely accepted that AI technologies can be used to scale cyberattacks via various forms of probing and automation. Multiple research and gaming efforts within cybersecurity communities have demonstrated the power using AI methods to attack computing systems. This area of work is referred to as *offensive AI*.[29,30]

4.1 Approaches to offensive AI

Offensive AI methods will likely be taken up as tools of the trade for powering and scaling cyberattacks. We must prepare ourselves for adversaries who will exploit AI methods to increase the coverage of attacks, the speed of attacks, and the likelihood of successful outcomes. We expect that uses of AI in cyberattacks will start with sophisticated actors but will rapidly expand to the broader ecosystem via increasing levels of cooperation and commercialization of their tools.[31]

Basic automation. Just as defenders use AI to automate their processes, so too can adversaries introduce efficiencies and efficacies for their own benefit. Automating attacks using basic pre-programmed logic is not new in cybersecurity. Many malware and ransomware variants over the last five years have used relatively simple sets of logical rules to recognize and adapt to operating environments. For example, it appears that attacking software has checked time zones to adapt to local working hours and customized behavior in a variety of ways to avoid detection or take tailored actions to adapt to the target computing environment.[32,33] On another front, automated bots have begun to proliferate on social media platforms.[34] These are all rudimentary forms of AI that encode and harness an attacker's expert knowledge. However, substantial improvements in AI technology make plausible malicious software that is much more adaptive, stealthy, and intrusive.[35]

Authentication-based attacks. AI methods can be employed in authentication-based attacks, where, for example, recently developed AI methods can be used to generate synthetic voiceprints to gain access through an authentication system. Compelling demonstrations of voice impersonations to fool an authentication system were presented during the Capture the Flag (CTF) cybersecurity competition at the 2018 DEF CON meeting.[36]

AI-powered social engineering. Human perception and psychology are weak links in cyber-defense. AI can be used to exploit this persistent vulnerability. We have seen the rise of uses of AI for *social engineering*, aiming the power of machine learning at influencing the actions of people to perform tasks that are not in their

[26] Cybersecurity Incidents (opm.gov)
[27] Russian Interference in 2016 U.S. Elections - FBI
[28] Characterizing networks of propaganda on twitter: a case study
[29] https://arxiv.org/pdf/2106.15764.pdf
[30] B. Buchanan, J. Bansemer, D. Cary, et al., Automating Cyber Attacks: Hype and Reality, Center for Security and Emerging Technology, November 2020. https://cset.georgetown.edu/wp-content/uploads/CSET-Automating-Cyber-Attacks.pdf
[31] How cyberattacks are changing according to new Microsoft Digital Defense Report
[32] Intelligence, FireEye Threat. "HAMMERTOSS: Stealthy tactics define a Russian cyber threat group." *FireEye, Milpitas, CA* (2015).
[33] Virtualization/Sandbox Evasion, Technique T1497 - Enterprise | MITRE ATT&CK®
[34] https://www.jmir.org/2021/5/e26933/
[35] See for example, see documentation of Deep Exploit, tools and demonstration showing the use of reinforcement learning to drive cyberattacks: https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit
[36] https://www.defcon.org/

interest. As an example, AI methods can be used to generate ultra-personalized phishing attacks capable of fooling even the most security conscious users. A striking 2018 study demonstrated how AI methods could be used to significantly raise the probability that end users would click on malevolent links in social media posts. The AI system learned from publicly available data including online profiles, connections, content of posts, and online activity of targeted individuals. Machine-learning was used to optimize the timing and content of messages with a goal of maximizing clickthrough rates—with significant results.[37] A 2021 study demonstrated that the language of emails could be crafted automatically with large-scale neural language models and that the AI-generated messages *were more successful than the human-written messages* by a significant margin.[38] In a related direction, Microsoft has tracked groups that use AI to craft convincing but fake social media profiles as lures.

4.2 AI-powered cyberattacks on the frontier

The need to prepare for more sophisticated offensive AI was highlighted in presentations at a National Academies of Sciences workshop on offensive AI that I co-organized in 2019. The workshop, sponsored by the Office of the Director of National Intelligence, led to a report available from the Academies.[39] The report includes discussion of the applications of AI methods across the cyber kill-chain, including the use of AI methods in social engineering, discovery of vulnerabilities, exploiting development and targeting, and malware adaptation, as well as in methods and tools that can be used to target vulnerabilities in AI-enabled systems, such as autonomous systems and controls used in civilian and military applications.

The cybersecurity research community has demonstrated the power of AI and other sophisticated computational methods in cyberattacks. Adversaries can harness AI to efficiently guess passwords, to attack industrial control systems without raising suspicions, and to create malware that evades detection or prevents inspection[40,41,42,43 ,44 ,45] AI-enabled bots can also automate network attacks and make it difficult to extinguish the

[37] J. Seymour and P. Tully, *Generative Models for Spear Phishing Posts on Social Media*, 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017. https://arxiv.org/abs/1802.05196

[38] https://www.wired.com/story/ai-phishing-emails/amp

[39] Implications of Artificial Intelligence for Cybersecurity: A Workshop, National Academy of Sciences, 2019. https://www.nationalacademies.org/our-work/implications-of-artificial-intelligence-for-cybersecurity-a-workshop

[40] Hey, My Malware Knows Physics! Attacking PLCs with Physical Model Aware Rootkit – NDSS Symposium (ndss-symposium.org)

[41] B. Hitaj, P. Gasti, G. Ateniese, F. Perez-Cruz, PassGAN: A Deep Learning Approach for Password Guessing, NeurIPS 2018 Workshop on Security in Machine Learning (SecML'18), December 2018. https://github.com/secml2018/secml2018.github.io/raw/master/PASSGAN_SECML2018.pdf

[42] S. Datta, DeepObfusCode: Source Code Obfuscation through Sequence-to-Sequence Networks In: Arai, K. (eds) Intelligent Computing. Lecture Notes in Networks and Systems, vol 284. Springer, Cham. https://doi.org/10.1007/978-3-030-80126-7_45, July 2021.

[43] J. Li, L. Zhou, H. Li, L. Yan and H. Zhu, "Dynamic Traffic Feature Camouflaging via Generative Adversarial Networks," *2019 IEEE Conference on Communications and Network Security (CNS)*, 2019, pp. 268-276, doi: 10.1109/CNS.2019.8802772. https://ieeexplore.ieee.org/abstract/document/8802772

[44] C. Novo, R. Morla, Flow-Based Detection and Proxy-Based Evasion of Encrypted Malware C2 Traffic, Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security 2020, https://doi.org/10.1145/3411508.3421379.

[45] D. Han *et al*., "Evaluating and Improving Adversarial Robustness of Machine Learning-Based Network Intrusion Detectors," in *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2632-2647, Aug. 2021, https://ieeexplore.ieee.org/abstract/document/9448103

attacker's command and control channels.[46] In another direction, a competitor demonstrated at a DARPA Cyber Grand Challenge exercise in 2016[47] how machine learning could be used to learn how to generate "chaff" traffic, decoy patterns of online activity that resemble the distribution of events seen in real attacks for distraction and cover-up of actual attack strategies.[48]

It is safe to assume that AI will improve the success, impact, and scope of the full breadth of threats present today. AI will also introduce new challenges, including special cyber vulnerabilities introduced with general uses of AI components and applications, which create new apertures for adversaries to exploit.

Recommendations: Prepare for malicious uses of AI to perform cyberattacks

- Raise DoD and other Federal agency awareness of the threat of AI-powered cyberattacks and directions with defenses against them, including detecting and thwarting new forms of automation and scaling.
- DoD should deeply engage with the cybersecurity community, participate in R&D and competitions on AI-enhanced cyberattacks and continue to learn from frontier advances, findings, and proposed mitigations.
- Increase R&D funding for exploring challenges and opportunities at the convergence of AI and cybersecurity. Consider the establishment of federally funded R&D centers of excellence in cybersecurity. Execute on the NSCAI recommendation to invest in DARPA to facilitate greater research on AI-enabled cyber defenses.[49]
- Formalize and make more efficient cross-sector networks for sharing updates on evolving technologies, data, attack vectors, and attacks.

5. Special vulnerabilities of AI systems

The power and growing reliance on AI generates a perfect storm for a new type of cyber-vulnerability: *attacks targeted directly at AI systems and components*. With attention focused on developing and integrating AI capabilities into applications and workflows, the security of AI systems themselves is often overlooked. However, adversaries see the rise of new AI attack surfaces growing in diversity and ubiquity and will no doubt be pursuing vulnerabilities. Attacks on AI systems can come in the form of *traditional vulnerabilities*, via *basic manipulations and probes*, and via a new, troubling category: *adversarial AI*.

5.1 Attacks on AI Supply Chains

AI systems can be attacked via targeting traditional security weaknesses and software flaws, including attacks on the supply chain of AI systems, where malevolent actors gain access and manipulate insecure AI code and data. As an example, in 2021, a popular software platform used to build neural networks was found to have 201 traditional security vulnerabilities, such as memory corruption and code execution.[50] Researchers have demonstrated how adversaries could use existing cyberattack toolkits to attack core infrastructure of the

[46] A botnet-based command and control approach relying on swarm intelligence - ScienceDirect

[47] https://www.darpa.mil/program/cyber-grand-challenge

[48] R. Rivest, Chaffing and Winnowing: Confidentiality Without Encryption," CryptoBytes, 4(1):12-17, https://pdfs.semanticscholar.org/aaf3/7e0afa43f5b6168074dae 2bc0e695a9d1d1b.pdf

[49] https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf. page 279.

[50] https://www.cvedetails.com/product/53738/Google-Tensorflow.html

software running AI systems.[51] Multiple components of AI systems in the supply chain of AI systems can be modified or corrupted via traditional cyberattacks. As an example, data sets used to train AI systems are rarely under version control in the same way that source code is. Researchers from NYU found that most AI frameworks downloaded from a popular algorithm repository do not check the integrity of AI models, in contrast to the standards of practice with traditional software, where cryptographic verification of executables/libraries has been standard practice for well over a decade.[52]

5.2 Adversarial AI

Adversarial AI or *adversarial machine learning* methods harness more sophisticated AI techniques to attack AI systems. Several classes of adversarial AI have been identified, including *adversarial examples*, the use of basic policies or more sophisticated machine learning methods to fool AI systems with inputs that cause the systems to fail to function properly. A second type of attack is called *data poisoning*, where data used to train AI systems are "poisoned" with streams of data that inject erroneous or biased training data into data sets, changing the behavior or degrading the performance of AI systems.[53] A third type of attack, called *model stealing*, seeks to learn details about the underlying AI model used in an AI system.[54] A fourth category of attack, called *model inversion,* seeks to reconstruct the underlying private data that is used to train the target system.[55]

With adversarial examples, basic manipulations or more sophisticated application of AI methods are used to generate inputs that are custom-tailored to cause failures in targeted AI systems. Goals of these attacks include disruptive failures of automated message classifiers, perceptions of machine vision systems, and recognitions of the words in utterances by speech recognition systems.

As an example of basic manipulations of inputs, a group, alleged to be within the Chinese government, attempted to amplify propaganda on Uyghurs by bypassing Twitter's anti-spam algorithm via appending random characters at the end of tweets.[56] The approach was viewed as an attempt to mislead the algorithm into thinking each tweet was unique and legitimate. In another example, researchers from Skylight appended benign code from a gaming database to Wannacry ransomware to cause the machine-learning-based antivirus filter to classify the modified ransomware as benign.[57] In related work on the fragility of AI systems, researchers showed that simply rotating a scan of a skin lesion confuses a computer recognition system to classify the image as malignant.[58]

[51] Xiao, Qixue, et al. "Security risks in deep learning implementations." 2018 IEEE Security and privacy workshops (SPW). IEEE, 2018.

[52] Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. "Badnets: Identifying vulnerabilities in the machine learning model supply chain." arXiv preprint arXiv:1708.06733 (2017).

[53] Jagielski, Matthew, et al. "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning." *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018.

[54] Yu, Honggang, et al. "CloudLeak: Large-Scale Deep Learning Models Stealing Through Adversarial Examples." NDSS. 2020.

[55] Ziqi Yang, Ee-Chien Chang, Zhenkai Liang, Adversarial Neural Network Inversion via Auxiliary Knowledge Alignment, 2019

[56] https://www.nytimes.com/interactive/2021/06/22/technology/xinjiang-uyghurs-china-propaganda.html

[57] https://skylightcyber.com/2019/07/18/cylance-i-kill-you/

[58] Finlayson, Samuel G., et al. "Adversarial attacks on medical machine learning." Science 363.6433 (2019): 1287-1289.

In uses of AI to generate adversarial examples, researchers have demonstrated stunning examples of failures. In one approach, adversarial methods are used to inject patterns of pixels into images to change what an AI system sees. While the changes with AI inferences are dramatic*, the changes to the original images are not detectable by humans*. Sample demonstrations include the modification of a photo of a panda leading an AI system to misclassify the panda as a gibbon and changes to a stop sign to misclassify it as a yield sign.[59,60] Similar demonstrations have been done in the realm of speech recognition, with the injection of hidden acoustical patterns in speech that changes what a listening system hears.[61] Attacks leading to such misclassifications and malfunctions can be extremely costly, particularly in high-stakes domains like defense, transportation, healthcare, and industrial processes.

Challenges of adversarial AI and a set of recommendations are called out in the final report of the National Security Commission on AI (NSCAI).[62] I chaired the lines of effort on directions with developing and fielding trustworthy, responsible, and ethical AI applications, leading to chapters 7 and 8 of the report and the appendix on NSCAI's recommendations on key considerations for fielding AI systems that align with democratic values, civil liberties, and human rights.[63,64,65] Chapter 7 of the report covers rising concerns with adversarial AI, including the assessment that, "*The threat is not hypothetical: adversarial attacks are happening and already impacting commercial ML systems*." In support of this statement, over the last five years, the Microsoft cybersecurity team has seen an uptick in adversarial AI attacks.[66] I believe the trend will continue.

5.3 Efforts to Mitigate Adversarial AI

Pursuit of resistant systems. Computer science R&D has been underway on methods for making AI systems more resistant to adversarial machine learning attacks. One area of work centers on raising the level of robustness of systems to attacks with adversarial inputs as described above.[67,68] Approaches include special training procedures to include adversarial examples, validation of inputs to identify specific properties that can reveal signs of an attack and making changes to the overall approach to building models, and modifying the objective functions used in optimization procedures used to create the models so that more robust models are created. While the latter techniques and research directions behind them are promising, the challenges of

[59] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, ICLR 2015. https://arxiv.org/pdf/1412.6572.pdf

[60] N. Papernot, P. McDaniel, I. Goodfellow, et al., Practical Black-Box Attacks against Machine Learning, ASIA CCS '17, April 2017. https://dl.acm.org/doi/pdf/10.1145/3052973.3053009

[61] M. Alzantot, B. Balaji, M. Srivastava, Did you hear that? Adversarial Examples Against Automatic Speech Recognition, Conference on Neural Information Processing Systems, December 2017. https://arxiv.org/pdf/1801.00554.pdf

[62] https://www.nscai.gov/

[63] "Upholding Democratic Values: Privacy, Civil Liberties, and Civil Rights in Uses of AI for National Security," Chapter 8*, Report of the National Security Commission on AI*, March 2021. https://reports.nscai.gov/final-report/chapter-8/

[64] "Establishing Justified Confidence in AI Systems," Chapter 8, *Report of the National Security Commission on AI,* March 2021. https://reports.nscai.gov/final-report/chapter-7/

[65] E. Horvitz J. Young, R.G. Elluru, C. Howell, Key Considerations for the Responsible Development and Fielding of Artificial Intelligence, National Security Commission on AI, April 2021. https://arxiv.org/ftp/arxiv/papers/2108/2108.12289.pdf

[66] Kumar, Ram Shankar Siva, et al. Adversarial machine learning-industry perspectives. *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2020.

[67] https://cacm.acm.org/magazines/2018/7/229030-making-machine-learning-robust-against-adversarial-inputs/fulltext

[68] A. Madry, A. Makelov, L. Schmidt, et al. Towards deep learning models resistant to adversarial attacks, ICLR 2018. https://arxiv.org/pdf/1706.06083.pdf

adversarial examples persist, per the large space of inputs to machine learning procedures. Thus, it is important to continue to invest in R&D on adversarial AI, to perform ongoing studies with red-teaming exercises, and to remain vigilant.

5.4 Tracking, Awareness, and Resources

Front-line awareness. Despite the opportunities that adversarial AI methods will provide to state and non-state actors for manipulating and disrupting critical AI systems and rising evidence of real-world attacks with adversarial AI, the idea of protecting AI systems from these attacks has been largely an afterthought. There is an urgency to be aware and to be ready to respond to adversarial AI threats, especially those used in critical areas such as defense. A Microsoft survey of 28 organizations in 2020 showed, despite the rise in attacks on AI systems, companies are still unaware of these kinds of intentional failures to AI systems and are massively underinvested in tools and processes to secure AI systems.[67] Ryan Fedasiuk, a noted researcher at Georgetown's Center for Security of Emerging Technology specializing in China's AI operations, notes that Chinese military officers have explicitly called out that the U.S. defenses are susceptible to data poisoning, and even so far as calling data integrity as "the Achilles' heel" of the U.S. joint all-domain command and control strategy.[69]

Resources and Engagement. Microsoft, along with MITRE and 16 other organizations created the *Adversarial ML Threat Matrix* to catalog threats to AI systems.[70] The content includes documentation of case studies where attacks have been made on commercial AI systems. For engineers and policymakers, Microsoft, in collaboration with Berkman Klein Center at Harvard University, released a taxonomy of machine learning failure modes.[71] For security professionals, Microsoft has open-sourced Counterfit, its own tool for assessing the posture of AI systems.[72] For the broader community of cybersecurity practitioners interested in AI and security, Microsoft hosts the annual Machine Learning Evasion Competition as a venue to exercise their muscle in attacking and securing AI systems.[73] Within the Federal government, the DoD has listed safety and security of AI systems in its core AI principles.[74] And there is encouraging activity by NIST on an AI Risk Assessment Framework to address multiple dimensions of AI systems, including robustness and security.[75]

Recommendations: Raise awareness and address vulnerabilities of AI systems

- Secure engineering supply chains for Federal AI systems, including use of state-of-the-art integrity checking for data, executables, libraries, and platforms used to construct AI systems; ensure that a security development lifecycle approach is in place for sensitive code and data.
- Require security reviews of AI engineering projects at DoD and other Federal AI agencies.
- Bring AI development and cybersecurity teams together to establish best practices and review programs.

[69] https://breakingdefense.com/2021/11/china-invests-in-artificial-intelligence-to-counter-us-joint-warfighting-concept-records/

[70] https://atlas.mitre.org/

[71] https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

[72] https://github.com/Azure/counterfit/

[73] https://mlsec.io/

[74] https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

[75] https://www.nist.gov/itl/ai-risk-management-framework

- Raise DoD awareness of challenges of adversarial AI and consider the vulnerabilities of AI systems and components.
- Pursue the use of robust machine learning algorithms to bolster resilience of systems in the face of adversarial examples.
- Develop training programs to raise awareness of cybersecurity and AI engineering workforce on security vulnerabilities of AI systems and components, risk of attacks with adversarial AI methods, and means for reducing risks.
- Invest in R&D on trustworthy, robust, and secure AI systems.

6. AI in Malign Information Operations

Advances in machine learning and graphics have boosted the abilities of state and non-state actors to fabricate and distribute high-fidelity audiovisual content, referred to as *synthetic media* and *deepfakes.* AI technologies for generating deepfakes can now fabricate content that is indistinguishable from real-world people, scenes, and events, threatening national security. Advances that could only be found with the walls of computer science laboratories or in demonstrations that surprised attendees at academic AI conferences several years ago are now widely available in tools that create audio and audiovisual content that can be used to drive disinformation campaigns.

6.1 Challenges of Synthetic Media

Advances in the capabilities of generative AI methods to synthesize a variety of signals, including high-fidelity audiovisual imagery, have significance for cybersecurity. When personalized, the use of AI to generate deepfakes can raise the effectiveness of social-engineering operations (discussed above) in persuading end-users to provide adversaries with access to systems and information.

On a larger scale, the generative power of AI methods and synthetic media have important implications for defense and national security. The methods can be used by adversaries to generate believable statements from world leaders and commanders, to fabricate persuasive false-flag operations, and to generate fake news events. A recent demonstration includes the multiple examples of manipulated and more sophisticated deepfakes that have come to the fore over the course of the Russian attack on Ukraine. This includes a video of President Volodymyr Zelenskyy appearing to call for surrender.[76]

The proliferation of synthetic media has had another concerning effect: malevolent actors have labeled real events as "fake," taking advantage of new forms of deniability coming with the loss of credibility in the deepfake era. Video and photo evidence, such as imagery of atrocities, are being called fake. Known as the "liar's dividend", the proliferation of synthetic media emboldens people to claim real media as "fake," and creates plausible deniability for their actions.[77]

We can expect synthetic media and its deployment to continue grow in sophistication over time, including the persuasive interleaving of deepfakes with unfolding events in the world and real-time synthesis of deepfakes. Real-time generations could be employed to create compelling, interactive imposters (e.g.,

[76] See: https://www.youtube.com/watch?v=X17yrEV5sl4

[77] The Liar's Dividend: The Impact of Deepfakes and Fake News on Politician Support and Trust in Media | GVU Center (gatech.edu)

appearing in teleconferences and guided by a human controller) that appear to have natural head pose, facial expressions, and utterances. Looking further out, we may have to face the challenge of synthetic fabrications of people that can engage autonomously in persuasive real-time conversations over audio and visual channels.

6.2 Direction: Digital Content Provenance

A promising approach to countering the threat of synthetic media can be found in a recent advance, named *digital content provenance* technology. Digital content provenance leverages cryptography and database technologies to certify *the source and history of edits* (the provenance) of any digital media. This can provide "glass-to-glass" certification of content, from the photons hitting the light-sensitive surfaces of cameras to the light emitted from the pixels of displays, for secure workloads. We pursued an early vision and technical methods for enabling end-to-end tamper-proof certification of media provenance in a cross-team effort at Microsoft.[78,79] The aspirational project was motivated by our assessment that, in the long-term, neither humans nor AI methods would be able to reliably distinguish fact from AI-generated fictions—and that we must prepare with urgency for the expected trajectory of increasingly realistic and persuasive deepfakes.

After taking the vision to reality with technical details and the implementation of prototype technologies for certifying the provenance of audiovisual content, we worked to build and contribute to cross-industry partnerships, including Project Origin, the Content Authenticity Initiative (CAI), and the Coalition for Content Provenance and Authenticity (C2PA), a multistakeholder coalition of industry and civil society organizations. [80,81,82,83] In January 2022, C2PA released a specification of a standard that enables the interoperability of digital content provenance systems.[84,85] Commercial production tools are now becoming available in accordance with the C2PA standard that enable authors and broadcasters to assure viewers about the originating source and history of edits to photo and audiovisual media.

The final report of the NSCAI recommends that digital content provenance technologies should be pursued to mitigate the rising challenge of synthetic media. In Congress, the bipartisan Deepfake Task Force Act (S. 2559) proposes the establishment of the National Deepfake and Digital Provenance Task Force.[86] Microsoft and its media provenance collaborators encourage Congress to move forward with standing-up a task force to help identify and address the challenges of synthetic media and we would welcome the opportunity to provide assistance and input into the work.

[78] P. England, H.S. Malvar, E. Horvitz, et al. AMP: Authentication of Media via Provenance, ACM Multimedia Systems 2021. https://dl.acm.org/doi/abs/10.1145/3458305.3459599

[79] E. Horvitz, A promising step forward on disinformation, *Microsoft on the Issues*, February 2021. https://blogs.microsoft.com/on-the-issues/2021/02/22/deepfakes-disinformation-c2pa-origin-cai/

[80] Project Origin, https://www.originproject.info/about

[81] J. Aythora, et al. Multi-stakeholder Media Provenance Management to Counter Synthetic Media Risks in News Publishing, International Broadcasting Convention 2020 (IBC 2020), Amsterdam, NL 2020 https://www.ibc.org/download?ac=14528

[82] Content Authenticity Initiative, https://contentauthenticity.org/

[83] Coalition for Content Provenance and Authenticity (C2PA), https://c2pa.org/

[84] C2PA Releases Specification of World's First Industry Standard for Content Provenance, Coalition for Content Provenance and Authenticity, January 26, 2022, https://c2pa.org/post/release_1_pr/

[85] https://erichorvitz.com/A_Milestone_Reached_Content_Provenance.htm

[86] Deepfake Task Force Act, S. 2559, 117th Congress, https://www.congress.gov/bill/117th-congress/senate-bill/2559/text

Recommendations: Defend against malign information operations

- Enact the Deepfake Task Force Act.
- Promote uses of digital media provenance for news and communications in defense and civilian settings.
- Adopt pipelines and standards for certifying digital content provenance of signals, communications, and news at DoD and other Federal agencies, prioritized by risk and disruptiveness of fabricated content.
- Review potential disruptions that malign information campaigns could have on DoD planning, decision making, and coordination based on manipulative uses of sophisticated fabrications of audiovisual and other signals, spanning traditional Signals Intelligence (SIGINT) pipelines, real-time defense communications, and public news and media.
- Invest in R&D on methods aimed at detection, attribution, and disruption of AI-enabled malign information campaigns.

Summary

I have covered in my testimony status, trends, examples, and directions ahead with rising opportunities and challenges at the intersection of AI and cybersecurity. AI technologies will continue to be critically important for enhancing cybersecurity in military and civilian applications. AI methods are already qualitatively changing the game in cyber defense. Technical advances in AI have helped in numerous ways, spanning our core abilities to prevent, detect, and respond to attacks—including attacks that have never been seen before. AI innovations are amplifying and extending the capabilities of security teams across the country.

On the other side, state and non-state actors are beginning to leverage AI in numerous ways. They will draw new powers from fast-paced advances in AI and will continue to add new tools to their armamentarium. We need to double down with our attention and investments on threats and opportunities at the convergence of AI and cybersecurity. Significant investments in workforce training, monitoring, engineering, and core R&D will be needed to understand, develop, and operationalize defenses for the breadth of risks we can expect with AI-powered cyberattacks. The threats include new kinds of attacks, including those aimed squarely at AI systems. The DoD, federal and state agencies, and the nation need to stay vigilant and stay ahead of malevolent adversaries. This will take more investment and commitment to fundamental research and engineering on AI and cybersecurity, and in building and nurturing our cybersecurity workforce so our teams can be more effective today—and well-prepared for the future.

Thank you for the opportunity to testify. I look forward to answering your questions.