THE IMPLICATIONS OF ARTIFICIAL INTELLIGENCE IN CYBERSECURITY

SHIFTING THE OFFENSE-DEFENSE BALANCE

JENNIFER TANG
TIFFANY SAADE
STEVE KELLY

OCTOBER 2024

IST Institute for SECURITY + TECHNOLOGY

The Implications of Artificial Intelligence in Cybersecurity:
Shifting the Offense-Defense Balance

October 2024
Authors: Jennifer Tang, Tiffany Saade, Steve Kelly

Design: Lillian Ilsley-Greene

About the Institute for Security and Technology

Uniting technology and policy leaders to create actionable solutions to emerging security challenges

Technology has the potential to unlock greater knowledge, enhance our collective capabilities, and create new opportunities for growth and innovation. However, insecure, negligent, or exploitative technological advancements can threaten global security and stability. Anticipating these issues and guiding the development of trustworthy technology is essential to preserve what we all value.

The Institute for Security and Technology (IST), the 501(c)(3) critical action think tank, stands at the forefront of this imperative, uniting policymakers, technology experts, and industry leaders to identify and translate discourse into impact. We take collaborative action to advance national security and global stability through technology built on trust, guiding businesses and governments with hands-on expertise, in-depth analysis, and a global network.

We work across three analytical pillars: the **Future of Digital Security**, examining the systemic security risks of societal dependence on digital technologies; **Geopolitics of Technology**, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; and **Innovation and Catastrophic Risk**, providing deep technical and analytical expertise on technology-derived existential threats to society.

Learn more: https://securityandtechnology.org/.

Acknowledgments

We are immensely grateful for the generous support of Google.org, whose funding supports IST's AI in cybersecurity work.

We would also like to extend our sincere thanks to the numerous survey participants and industry experts who provided invaluable insights and feedback on the state of AI integration in the cybersecurity landscape. Their expertise and willingness to contribute have been critical to the depth and rigor of this paper.

Finally, and perhaps most importantly, our heartfelt gratitude goes out to those who are at the cutting edge of this work, living it and striving everyday to protect our nation's networks, economic security, and who face long odds every minute of the day. You deserve all the support we can provide to turn the advantage in your direction.

Table of Contents

Executive Summary

Cutting edge advances in artificial intelligence (AI) are taking the world by storm, driven by a massive surge of investment, countless new start-ups, and regular technological breakthroughs. AI presents key opportunities within cybersecurity, but concerns remain regarding the ways malicious actors might also use the technology. In this study, the Institute for Security and Technology (IST) seeks to paint a comprehensive picture of the state of play—cutting through vagaries and product marketing hype, providing our outlook for the near future, and most importantly, suggesting ways in which the case for optimism can be realized.

The report concludes that in the near term, AI offers a significant advantage to cyber defenders, particularly those who can capitalize on their "home field" advantage and first-mover status. However, sophisticated threat actors are also leveraging AI to enhance their capabilities, making continued investment and innovation in AI-enabled cyber defense crucial. At this time of writing, AI is not yet unlocking novel capabilities or outcomes, but instead represents a significant leap in speed, scale, and completeness.

This work is the foundation of a broader IST project to better understand which areas of cybersecurity require the greatest collective focus and alignment—for example, greater opportunities for accelerating threat intelligence collection and response, democratized tools for automating defenses, and/or developing the means for scaling security across disparate platforms—and to design a set of actionable technical and policy recommendations in pursuit of a secure, sustainable digital ecosystem.

Findings & Recommendations

Premise #1: AI has revolutionized content analysis, radically assisting defenders and offenders alike.

AI significantly assists both defenders and offenders in analyzing vast amounts of data, generating actionable insights at speed and scale.

→ **Recommendation A**: Protect sensitive data from malicious AI-enabled content analysis.

Implement robust cybersecurity practices, including data encryption, least privilege access, and multi-factor authentication. Carefully manage AI copilots to prevent unauthorized access to sensitive information. Ensure information security controls are not undermined by AI features within the organization.

Premise #2: AI challenges user authentication and human-to-human interactions.

AI-powered deepfakes and sophisticated phishing threaten traditional authentication methods and social trust.

→ **Recommendation B**: Supplement watermarking with alternative deepfake detection approaches.

Implement a multi-layered approach to content authenticity. Use watermarking alongside content provenance techniques and advanced detection technologies. Invest in platform-level detection capabilities to flag potentially deceptive content. Encourage the development and adoption of open standards for digital content verification.

→ **Recommendation C**: Modernize authentication approaches to account for AI.

Adopt physical authentication solutions leveraging public key cryptography, such as FIDO2 standards. Implement phishing-resistant multi-factor authentication. Explore the use of mobile driver's licenses (mDLs) for robust online identity verification. Develop human-based recovery protocols for contingencies to re-establish trust.

→ **Recommendation D**: Educate society to navigate the challenges brought by AI deepfakes.

Launch comprehensive awareness campaigns about AI-driven deception. Promote critical thinking and mindful content consumption. Invest in user-friendly tools for content verification. Encourage platforms to implement adaptive detection tools to flag AI-generated deceptive content for users.

Premise #3: AI will make software vastly more secure.

AI enhances code writing, reviewing, and vulnerability detection, but can also introduce new risks.

→ **Recommendation E**: Optimize both human and AI resources to achieve efficiency and software quality.

Integrate AI code assistants thoughtfully, balancing automation with human expertise. Implement rigorous quality assurance processes for AI-generated code. Foster a culture of critical assessment and continuous learning among developers. Use AI for lower-risk tasks initially, gradually increasing complexity as confidence grows.

Premise #4: AI will revolutionize security operations.

AI streamlines and enhances various aspects of cybersecurity operations, acting as a significant force multiplier.

→ **Recommendation F**: Integrate AI into security operations workflows, but protect your models.

Carefully integrate AI models into existing security frameworks. Maintain human oversight for nuanced decision-making. Implement comprehensive protection measures for AI models, including robust data governance and lifecycle management. Conduct regular red team exercises to identify and address vulnerabilities in AI systems.

Premise #5: AI is supercharging adversarial reconnaissance and target identification.

AI enables threat actors to conduct more efficient and effective reconnaissance of targeted networks.

→ **Recommendation G**: Minimize external attack surface; for critical systems, strive for invisibility.

Reduce public internet-exposed assets, implement effective network segmentation and strict controls between information technology (IT) and operational technology (OT) environments. Adopt zero trust architecture principles. Consider network obfuscation techniques to minimize the discoverable attack surface.

Introduction

Cutting edge advances in artificial intelligence (AI) are taking the world by storm, driven by broad accessibility, a massive surge of investment, new start-ups, regular research and technological breakthroughs, and an insatiable appetite by actors of all stripes to realize its potential across a myriad of domains. As automation becomes increasingly embedded in essential government functions and life-sustaining infrastructure services—such as food supply chains, healthcare systems, and public safety mechanisms—global leaders are recognizing that it is more imperative than ever to harness AI for protection, better understand how it may be weaponized by bad actors, and address digital security threats that contribute to systemic global risk.

The cybersecurity field was an early adopter of Machine Learning (ML) capabilities, empowering actors on both defense and offense alike, but the more recent entrée of Large Language Models (LLMs) and generative AI raises new questions on where the advantage lies today and in the future. Central to this debate is the "defender's dilemma," a common truism that posits, other things being equal, that a cyber attacker has the advantage over a cyber defender. (This truism is notable, as it challenges the oft-accepted view that attackers must vastly outnumber well entrenched defenders in the context of traditional military battle.) And given the complexity of managing a range of threats and vulnerabilities, the defender must always be right, while the attacker need only be right once to gain entry. While not everyone buys into this logic, any difference of opinion takes nothing away from the sustained yeoman's work required to successfully manage cyber risk in any large and complex organization.[1] The defender's task is hard, and the attacker's barrier to entry is often low.

An optimistic view is emerging among industry heavyweights and other responsible actors that recent leaps in AI technology bring the potential to dramatically alter the offense-defense balance by drawing on the home field advantage. This perspective highlights AI's potential to fortify defense mechanisms and outmaneuver adversaries, offering a "once-in-a-generation moment to change the dynamics of cyberspace for the better."[2]

In this study, the Institute for Security and Technology (IST) seeks to paint a comprehensive picture of the state of play—cutting through vagaries and product marketing hype, providing our outlook for the near future, and most importantly, suggesting ways in which the case for optimism can be realized.

1 David J. Blanco, "Cybersecurity teams, beware: The defender's dilemma is a lie," *TechCrunch*, February 7, 2023, https://techcrunch.com/2023/02/07/cybersecurity-teams-beware-the-defenders-dilemma-is-a-lie/.

2 "Secure, Empower, Advance: How AI Can Reverse the Defender's Dilemma," Google, February 2024, https://services.google.com/fh/files/misc/how-ai-can-reverse-defenders-dilemma.pdf.

This work is the foundation of a broader IST project to better understand which areas of cybersecurity require the greatest collective focus and alignment—for example, greater opportunities for accelerating threat intelligence collection and response, democratized tools for automating defenses, and/or developing the means for scaling security across disparate platforms—and to design a set of actionable technical and policy recommendations in pursuit of a secure, sustainable digital ecosystem.

Methodology

The authors leveraged recent literature and scholarship, including the Aspen Institute's report, "Envisioning Cyber Futures with A.I." and IST's recently published report, "A Lifecycle Approach to AI Risk Reduction: Tackling the Risk of Malicious Use Amid Implications of Openness" to set the scene regarding AI's practical integration into offensive and defensive cyber capabilities.[3,4]

To validate and augment this review, IST conducted a series of targeted surveys and interviews with industry incumbents, startups, consultancies, and threat researchers to capture current insights into how organizations and practitioners are currently engaging with or integrating AI technologies, the evolving impact of these tools on the threat landscape, and their forecasts for the future. Under the Chatham House rule, which allows use of the responses but without attribution, the survey posed four core questions:

- How is your organization currently using AI technologies to enhance cybersecurity for itself or client organizations?
    - What AI-specific or AI-enabled tools, techniques, or procedures are you making use of?
    - Has your approach measurably changed over the past couple of years (pre-AI mania vs. post; ML vs. LLMs of this year)? If so, how?
- From your organization's first-hand vantage point, are you seeing changes in the cyber threat landscape as a result of these technologies being used by malicious actors?
    - If yes, in what way?
- What is your outlook regarding the use of AI in cyber defenses (include what, when, and to what extent)?
- What is your outlook regarding the use of AI in cyber offense by bad actors (include what, when, and to what extent)?

3 "Envisioning Cyber Futures with A.I.," Aspen Institute, January 9, 2024, https://www.aspendigital.org/report/cyber-futures-with-ai/.

4 Louie Kangeter, "A Lifecycle Approach to AI Risk Reduction: Tackling the Risk of Malicious Use Amid Implications of Openness," Institute for Security and Technology, June 2024, https://securityandtechnology.org/virtual-library/reports/a-lifecycle-approach-to-ai-risk-reduction/.

The initial literature review informed the interviews, which generated new insights that inspired additional research, both in the form of additional written sources and stakeholder interviews. The authors' attendance at AI- and cyber-relevant meetings and conferences also piqued awareness and interest in additional ideas, sources, and use cases. This iterative approach led to the findings summarized below.

Analysis and Key Findings

The analysis that follows highlights five conclusions, termed "premises," that reflect both the current state of play and outlook for AI reshaping the cybersecurity landscape. Note, the tense of each premise suggests its temporal framing.

Premise #1: AI has revolutionized content analysis, significantly assisting defenders and offenders alike.

In recent months, users of the internet search services Google Search and Microsoft Bing have been treated to a new AI-enabled experience: the search overview or summary. This move brings to the mainstream functionality launched two years earlier by the conversational search engine Perplexity AI and marks a sea change in how humans will search, access, and consume knowledge and information. When posed a question, all three services generate cogent answers presented in clear narrative form, often drawing on numerous sources. For users not accustomed to this feature, it is startlingly useful, particularly when compared to the decades-long practice of scrolling through an endless list of search hits to eventually find an answer. Furthermore, the rise of "multimodal"[5] AI models have enabled both the processing and generation of content across different forms of media, such as text, images, and audio.

The technology behind this innovation can also be used by both good and bad actors to review, summarize, and interpret vast volumes of information that is not on the public internet. For example, an organization can use an AI search copilot to make internal records and knowledge bases more readily available and useful to employees, answering natural language questions such as, "What is the company's policy on mileage reimbursement?" or "Were any dogs present at the company's most recent retreat?" At the same time, a foreign intelligence service or ransomware actor can use this same technology, informed by a set of intelligence

5 A multimodal model integrates and processes information from multiple types of data ("modalities"). Unlike unimodal models, which focus on a single type of input (e.g., text, image, audio), multimodal models can synthesize information from different sources simultaneously, leading to improved performance in various applications such as content generation, translation, and interaction-based systems.

requirements or objectives, to translate and review stolen information to determine its meaning, value, and how it—and its personnel—might be further exploited.[6]

AI POWERING THE CYBER THREAT INTELLIGENCE CYCLE

The application of AI in cyber threat intelligence is rapidly transforming how organizations detect, analyze, and respond to emerging threats throughout the threat intelligence cycle. With vast amounts of data generated across networks, AI tools help sift through noise, identify anomalies, and predict potential attacks before they materialize. From automating threat hunting to enriching threat feeds with real-time insights, AI is evolving from a supplementary tool to a valuable component in defending against cyber adversaries. For example, by tracking notable threat actor profiles and cyber attacks by region and industry, predictive AI models can analyze "vast amounts of data to identify indicators of compromise (IOCs) or common TTPs used by attackers."[7] This capability could incentivize organizations to take a more proactive approach to both current and emerging cyber threats by gathering real-time data on the threat landscape and leveraging insights from past threat intelligence reports.

A Google Cloud report identified key emerging trends from the use of LLMs in threat intelligence, notably the enhancement of threat intelligence capabilities by integrating data points from different sources.[8] They found that LLMs have increased the coverage of digital threats in various languages, which helps more effectively identify and track bad actors across the globe. LLMs have also improved the personalization of threat intelligence outputs based on the specific threat exposure of a given organization, as well as the "actionability" of threat intelligence. In fact, some industry organizations have been combining their threat intelligence with security operations, which includes "using AI-based models to curate and prioritize indicators of compromise" in order to identify potenital breaches early on.[9]

> *"The most dangerous LLM cyber capability is summarization"*
>
> - Gabe Bernadett-Shapiro, AI security researcher
> X thread, August 11, 2024

Backed by LLMs, key phases of the threat intelligence cycle have already been augmented by AI. At the collection stage, data from the dark web, discussion forums, messaging services, and everything in between is aggregated with greater speed and accuracy, offering defenders

6 Mariami Tkeshelashvili and Tiffany Saade, "Decrypting Iran's AI-Enhanced Operations in Cyberspace," Institute for Security and Technology, September 26, 2024, https://securityandtechnology.org/blog/decrypting-irans-ai-enhanced-operations-in-cyberspace/.

7 Tiffany Saade, "Artificial Intelligence for Cyber Resilience," Resilience, September 25, 2024, https://www.cyberresilience.com/threatonomics/artificial-intelligence-for-cyber-resilience/.

8 Vijay Ganti and Scott Coull, "Introducing AI-powered insights in Threat Intelligence," Google Cloud, April 24, 2023, https://cloud.google.com/blog/products/identity-security/rsa-introducing-ai-powered-insights-threat-intelligence.

9 Ganti and Coull, "Introducing AI-powered insights."

actionable threat data.[10] Equally as important has been LLMs' ability to better "categorize and annotate the extracted data with valuable enrichments, thereby saving analysts hours of manual examination."[11] This enhanced data processing not only accelerates the initial stages of threat intelligence, but creates a feedback loop for continuous improvement in the models' ability to attribute threats more precisely. Analysts are also able to leverage AI for content analysis and classification by triaging the most relevant pieces of information within the pool they have collected. Automating this capability would eventually "speed up and scale [the analysts'] ability to put the puzzle pieces together."[12]

ADVERSARIAL USE OF AI CONTENT ANALYSIS

A cybersecurity company, Intel471, noted in a recent blog post that they "observed a threat actor claim to use Meta's Llama AI to search through breach data," which is notable since ransomware groups sometimes "exfiltrate terabytes of data, most of which may be mundane, so isolating sensitive data could be a viable AI use case."[13] Along the same lines, a prominent AI research scientist, Gabe Bernadett-Shapiro, recently asserted that "the most dangerous LLM cyber capability is summarization" for its ability to examine a pool of data and answer, "[W]hy do I care about this?" He also shared a use case in which an LLM might interpret a file structure and predict where high-value files are stored, such as cloud service access keys or a backup file. Such capabilities can help a bad actor who happens upon an unpatched system understand "Where [am] I? What does this system do? What does it have access to? Is it related to my interests?"[14] Given these developments, it is clear that the integration of AI into cybercrime strategies is not only imminent, but also poses a notable threat to cyber defenders.

This report joins those views. Widespread cyber attacks that simultaneously impact hundreds or even thousands of victims are increasingly common, challenging intruders' ability to quickly prioritize and exploit systems before the campaign is discovered and addressed. For a ransomware actor with untold victimization opportunities who is drowning in stolen data, it will no doubt revolutionize their ability to pick higher-value targets, extort more victims simultaneously, and significantly increase ransom demand payments.

In addition to almost every ransomware scheme, actors behind the most successful supply chain-based cyber attack campaigns—including those impacting SolarWinds, Kaseya,

10 Rey Bango, "How AI Can Improve Threat Intelligence Gathering and Usage," Microsoft, November 10, 2023, https://techcommunity.microsoft.com/t5/educator-developer-blog/how-ai-can-improve-threat-intelligence-gathering-and-usage/ba-p/3975449.

11 Scott Coull and Jayce Nichols, "AI and the FIve Phases of the Threat Intelligence Lifecycle," Google Cloud, August 24, 2023, https://cloud.google.com/blog/topics/threat-intelligence/ai-five-phases-intelligence-lifecycle.

12 Coull and Nichols, "AI and the Five Phases."

13 "Cybercriminals and AI: Not Just Better Phishing," Intel471, June 12, 2024, https://intel471.com/blog/cybercriminals-and-ai-not-just-better-phishing.

14 Gabe Bernadett-Shapiro (@Gabeincognito), "Clickbait and the Apocalypse," X, August 11, 2024, https://x.com/Gabeincognito/status/1822747622580642263.

and Microsoft Exchange Server—would have significantly benefited from AI-enabled summarization capabilities. It is reasonable to conclude that sophisticated nation-state actors are, or will soon be, using them to their full advantage.

Recommendation A: Protect sensitive data from malicious AI-enabled content analysis.

The risks presented above involve a matter of scale and intensity, and not of kind. Protecting sensitive and confidential information has long been a key cybersecurity priority and challenge. Given AI's ability to mine large datasets—including software code, natural language text, images, audio, and video—the "security through obscurity" fallacy is less true than ever before; there is no hiding in the noise. Organizations must quickly orient to cybersecurity best practices, such as those articulated in the Center for Internet Security's 18 Critical Security Controls, and implement data encryption, the principle of "least privilege" in user access, and multi-factor authentication, to name a few.[15]

Additionally, organizations implementing AI copilots within their environment must ensure that such features do not undermine information security controls by making sensitive information discoverable by only users who have a need to know. Upon gaining unauthorized access to a victim organization's network, intruders may also be able to take advantage of the organization's AI copilot to find information or network resources they are seeking, accelerating their attack cycle. Therefore, when training AI models, it is crucial for organizations to understand the derivative use of these models to prevent unwanted data exposure.

Premise #2: AI challenges user authentication and human-to-human interactions.

What was once relegated to the realm of doomsday hypotheticals, the challenges to user authentication and human interaction posed by AI are becoming increasingly commonplace. The boundaries between legitimate and deceptive digital interactions are blurring, making traditional user authentication methods increasingly vulnerable. AI's ability to craft convincing synthetic identities, whether through deep fakes or sophisticated phishing emails, places existing Identity Access Management (IAM) systems under strain. The precision of AI-driven impersonations is so effective that they can even evade traditional detection methods. This

15 "The 18 CIS Critical Security Controls," Center for Internet Security, accessed August 22, 2024, https://www.cisecurity.org/controls/cis-controls-list.

increased stealth and plausible deniability heighten concerns about the reliability of existing authentication mechanisms.

Identity Access Management (IAM) systems

As organizations grapple with the challenges posed by AI to user authentication, some are already putting forth solutions to enhance their IAM systems. For example, IBM's Verify and Google's IAM Recommender provide actionable insights to optimize access controls, helping organizations manage permissions effectively in a landscape increasingly threatened by AI-driven impersonations or attacks.[16,17] By analyzing usage patterns and recommending adjustments and recourse, organizations are already seeking to bolster their defenses.

Human trust is put to the test by the proliferation of AI-enabled impersonation. Being human is simply not enough anymore; the core question of "who can we trust" in an ever-evolving threat landscape amplified by AI is difficult to answer. As malicious actors evolve the art of phishing and identity manipulation through AI generated content, it prompts a critical examination of how these technologies are amplifying bad actors' ability to deceive, evade, and breach, and what organizations can do about it.

DEEPFAKES FUELING IMPERSONATION ATTACKS

The rise of convincing deepfake technology poses a severe threat to traditional authentication systems that rely on visual or auditory cues for verification. Biometric authentication systems that use facial recognition or voice analysis have already been compromised by deepfake technology in several cases.[18] For example, one surveyed organization revealed an instance where authentication systems were duped into validating synthetically generated biometrics.[19] In another case, a major cybersecurity firm identified a sophisticated mobile Trojan, GoldPickaxe, aimed at iOS users. This malware, believed to originate from a Chinese cybercrime group, is capable of harvesting facial recognition data, collecting identity documents, and intercepting text messages.[20] The stolen data was then used, in conjunction with AI-powered face-editing tools, to create deepfakes that granted unauthorized access to

16 "Cloud identity and access management solutions," IBM, accessed August 22, 2024, https://www.ibm.com/products/verify-saas/cloud..

17 "Overview of role recommendations," Policy Intelligence, Google Cloud, accessed September 18, 2024, https://cloud.google.com/policy-intelligence/docs/role-recommendations-overview.

18 Andrey Polovinkin and Sharmine Low, "Face Off: Group-IB identifies first iOS trojan stealing facial recognition data," Group-IB, February 15, 2024, https://www.group-ib.com/blog/goldfactory-ios-trojan/.

19 "How AI Enables Hacking of Biometric Authentication Systems," Cyber Brain Academy, July 12, 2023, https://cyberbrainacademy.com/how-ai-enables-hacking-of-biometric-authentication-systems/.

20 Polovinkin and Low, "Face-Off."

victims' bank accounts and other fraud schemes.[21] Unsurprisingly, malicious deepfake is on the rise.

In once instance demonstrating the threat posed by deepfakes, a senior finance worker at a multinational corporation received an urgent message purportedly from the company's Chief Financial Officer asking him to join an urgent video conference call with several other employees. During the call, the worker believed he was interacting with his colleagues; unbeknownst to him, all of the participants were, in fact, AI-generated deepfake re-creations. The deception was so effective that any initial doubts were allayed, and under pressure, the worker remitted a payout of over $25 million. By the time the ruse was discovered, the funds were irretrievable.[22]

Social engineering, already a complex challenge to cybersecurity, is becoming even more formidable with the proliferation of AI-enabled deception. As demonstrated in the above vignette, deepfake technology leverages advanced AI algorithms to generate highly realistic images, videos, and audio, mimicking real individuals with startling accuracy. Bad actors have already impersonated executives in phishing schemes, creating false identities for deception, and fabricating evidence in legal and financial fraud, among other dangerous use cases. The ability to detect these tactics is only becoming more difficult.[23,24] Following the release of OpenAI's Sora, which employs a text to video model, a HarrisX survey showed 1,000 American respondents a combination of eight AI-generated videos andvideos created with traditional tools to test their ability to identifty AI-generated content.[25] The survey results revealed that "most US adults incorrectly guessed whether AI or a person had created five out of the eight videos they were shown."[26,27]

FROM BAIT TO BREACH: THE RISE OF AI-ENABLED PHISHING THREATS

Phishing—whether by email, text message, or telephone call—is simply a modern variation of the confidence game, whereby a perpetrator uses deceptive tactics to gain the victim's trust. This old-fashioned fraud technique is by no means waning; in fact, CISA last year reported that

21 Polovinkin and Low, "Face-Off."

22 Heather Chen and Kathleen Magramo, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'," *CNN*, February 4, 2024, https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

23 Chen and Magramo, "Finance worker pays out $25 million."

24 Daniele Lepido, "'I Need to Identify You': How One Question Saved Ferrari From a Deepfake Scam," *Bloomberg*, July 26, 2024, https://www.bloomberg.com/news/articles/2024-07-26/ferrari-narrowly-dodges-deepfake-scam-simulating-deal-hungry-ceo.

25 Nil Shah, "Why watermarking is just one part of combating AI deepfakes," *Variety*, March 21, 2024, https://variety.com/vip/watermarking-just-one-part-of-combating-ai-deepfakes-1235946602/.

26 "Video generation models as world simulators," OpenAI, February 15, 2024, https://openai.com/index/video-generation-models-as-world-simulators/.

27 Audrey Schomer, "Sora AI Videos Easily Confused with Real Footage in Survey Test," *Variety*, March 8, 2024, https://variety.com/vip/sora-ai-video-confusion-human-test-survey-1235933647/.

90 percent of successful cyberattacks begin with phishing.[28] And these observations likely do not fully reflect bad actors' recent and growing adoption of generative AI to create more convincing phishing content.

Current LLMs can craft emails with convincing detail, providing such granularity and language fluency that they appear legitimate. Victims are often lured into divulging sensitive information and inadvertently compromising their security, creating vulnerabilities that traditional defenses, which often rely on the detection of grammatical errors and semantic inconsistencies, are struggling to manage effectively.[29] In an experiment consisting of 4,600 participants, most found it difficult to differentiate AI-generated content from human-generated content, with accuracy rates no better than random chance, ranging from 50 to 52 percent.[30] Participants often resorted to flawed heuristics—cognitive shortcuts used to judge the authenticity of language. These heuristics include the use of first-person pronouns, family-related content, and grammar. However, these cues are ineffective because AI could mimic these features, making human judgment predictable and manipulable by AI systems.

Across the board, organizations surveyed by IST reported a notable rise in AI-driven phishing attacks. Several respondents highlighted how threat actors leverage LLMs to conduct detailed research on targets and produce error-free, targeted phishing content, enhancing the efficacy of spear-phishing operations. For instance, Microsoft observed Iranian state-affiliated threat actor, Crimson Sandstorm, interacting with ChatGPT to "generate various phishing emails, including one pretending to come from an international development agency and another attempting to lure prominent feminists to an attacker-built website on feminism."[31]

This trend shows no signs of slowing down. According to IBM's X-Force Threat Intelligence Index, the most significant shift observed in 2023 was a surge in cyber threats that targeted one's identity, rather than vulnerabilities.[32] Zscaler, an industry leader in cloud security, reported a 58 percent increase in AI-enabled phishing attacks in 2024, while SlashNext observed an over 4,000 percent increase in malicious emails and nearly 1,000 percent increase in credential phishing since late 2022 following the advent of ChatGPT.[33,34]

---

28 "Be cyber smart: Get your "Shields Up" Simple Steps for Safety Online," Cybersecurity and Infrastructure Security Agency (CISA), February 2023, https://www.cisa.gov/sites/default/files/2023-02/cisa_fact_sheet_4_things_cyber_english_508.pdf.

29 Rick Liu, Jessica Choi, and Eva Kwok, "Deepfakes and AI: How a 200 Million Scam Highlights the Importance of Cybersecurity Vigilance," FTI Consulting, February 23, 2024, https://fticybersecurity.com/2024-02/deepfakes-and-ai-how-a-200-million-scam-highlights-the-importance-of-cybersecurity-vigilance/.

30 Maurice Jakesch, Jeffrey Hancock and Mor Naaman, "Human heuristics for AI-generated language are flawed," *PNAS* 120, no. 11 (2023), https://www.pnas.org/doi/pdf/10.1073/pnas.2208839120.

31 Microsoft Threat Intelligence, "Staying ahead of threat actors in the age of AI," Microsoft, February 14, 2024, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

32 "X-Force Threat Intelligence Index 2024," IBM, February 2024, https://www.ibm.com/downloads/cas/L0GKXDWJ.

33 Deepan Desai and Rohit Hedge, "Phishing Attacks Rise 58% in the Year of AI: ThreatLabz 2024 Phishing Report," Zscaler, April 23, 2024, https://www.zscaler.com/blogs/security-research/phishing-attacks-rise-58-year-ai-threatlabz-2024-phishing-report.

34 "The State of Phishing: 2024 Mid-Year Assessment," SlashNext, April 10, 2024, https://slashnext.com/state-of-phishing-2023/.

Despite the rise of AI-enabled phishing campaigns, it is important to acknowledge that phishing-resistant authentication methods have been in widespread use for over a decade and effectively address many of these challenges. While there is legitimate concern regarding AI's capacity to generate near-perfect phishing content at scale, organizations can mitigate these risks through robust authentication protocols. Several industry giants have implemented advanced systems capable of identifying phishing emails with high certainty, analyzing both content and non-content factors, including domain registrations and email text.[35,36] Though these anti-phishing tools have been successfully deployed by industry heavyweights, many users may still have limited or no access to them (e.g., users might prefer other email apps that lack robust anti-phishing detection).

However, the rapid improvement of AI in generating near-perfect phishing content presents a significant challenge for human judgment. And as phishing messages become increasingly difficult to discern as malicious, the ability of individuals to critically assess the legitimacy of their communications and online media at-large is compromised. The key questions we need to ask ourselves when engaging with potentially malicious content—"Is this email trying to deceive me? Are these links safe? Is this really my boss?"—are becoming increasingly difficult to answer. This challenge is compounded when humans stop asking these questions because of our trust in (or overreliance on) technology.

THE UNRAVELING OF INDIVIDUAL AND SOCIAL TRUST

As AI-generated content becomes more sophisticated, the line between truth and fiction blurs, making it increasingly difficult to discern what is real, manipulated, or altogether fake. And the implications of this erosion are far-reaching, with both our interpersonal and social trust at risk of deterioration.

As societies scale, social trust—the expectation that individuals, groups, organizations, or institutions will behave in a particular way—is required to facilitate predictability and stability.[37] The information and communications technologies that underpin modern economic and social life are also built on a foundation of social trust. But if these mechanisms are consistently undermined by AI-powered deceptive content, society's trust in the institutions that once safeguarded our privacy, security, and well-being might begin to fracture.

35 "Advanced Protection Program," Google, accessed September 24, 2024, https://landing.google.com/advancedprotection/.

36 "Phishing protection and prevention solutions," Microsoft Security, accessed September 24, 2024, https://www.microsoft.com/en-ca/security/business/solutions/phishing.

37 Susan Verducci and Andreas Schröer, "Social Trust," in *International Encyclopedia of Civil Society,* eds. Helmut Anheier and Stefan Toepler (New York: Springer, 2010), https://doi.org/10.1007/978-0-387-93996-4_68.

Previous IST work explored how digital technologies affect human cognition, and what those effects might mean for democracy.[38] This work resulted in individual papers on memory, attention, reasoning, critical thinking, trust, and emotion. The culminating report observed that digital technologies influence the conduct and quality of democracy on three levels—cognitive, individual, and societal—by manipulating cognition directly or encouraging individuals to outsource their cognitive functions. The study team and working group members identified 12 techno-cognitive risks driven by gamification, information overload, immersive experiences, and a lack of friction in the user experience. These risks increase susceptibility to misinformation and drive political-ideological polarization.

IST's Generative Identity Initiative (GII) —which builds on the Digital Cognition for Democracy Initiative discussed above— seeks to address the complex questions around generative AI's impact on social identities, norms, and belonging.[39] As these personalized AI cyber issues become more and more prolific, the fallout could very likely be the atrophy of interpersonal trust, the confidence we have in each other as individuals. Once one begins to doubt the authenticity of a bank teller's identity or the voice of a loved one asking for help, the fundamental ability to engage in everyday social interactions and relationships may be compromised, potentially reshaping the very fabric of human interaction and social cohesion in the digital age. Fractured social and interpersonal trust undermines the effectiveness of long relied-upon security measures, leaving both individuals and institutions exposed to the many risks discussed throughout this report. The forthcoming proceedings of IST's Generative Identity Initiative will serve as a useful complement to this report.[40]

As established in the preceding subsections, AI is poised to supercharge these dynamics through a combination of data mining for individualized precision targeting and the creation of personalized inauthentic content. This is by no means dystopian fiction. A recent report from RAND highlights that researchers in China are developing an AI system designed for "precision cognitive attacks."[41] By leveraging advanced generative AI technologies, the system could process large amounts of data autonomously to analyze user preferences. It would then generate and deliver extremely tailored information, enabling attacks based on individualized user profiles and exploiting personalized cognitive weaknesses.

38 "Digital Cognition & Democracy Initiative," Institute for Security and Technology, accessed August 22, 2024, https://securityandtechnology.org/dcdi/.

39 "IST launches Generative Identity Initiative with support of Omidyar Network," Institute for Security and Technology, January 2024, https://securityandtechnology.org/blog/ist-launches-generative-identity-initiative-with-support-of-omidyar-network/.

40 "Generative Identity Initiative," Institute for Security and Technology, accessed September 1, 2024, https://securityandtechnology.org/generative-identity-initiative/.

41 Nathan Beauchamp-Mustafaga, "Exploring the Implications of Generative AI for Chinese Military Cyber-Enabled Influence Operations: Chinese Military Strategies, Capabilities, and Intent," RAND, February 1, 2024, https://www.rand.org/content/dam/rand/pubs/testimonies/CTA3100/CTA3191-1/RAND_CTA3191-1.pdf.

Recommendation B: Supplement watermarking with alternative deepfake detection approaches.

The deepfake problem can be tackled from two angles: marking authentic content, and marking inauthentic (i.e., AI-generated or AI-manipulated) content. Assuming for a moment that it is often possible to categorize content as authentic or inauthentic, it leaves a remaining category of content that is not immediately identifiable as either. This gives rise to the need for detection technologies to potentially close the gap. It should also be noted that marking content as authentic or inauthentic does not mean that it is trustworthy or not, only that its original source has been verified.

Watermarking aims to embed traceable markers within AI-generated media to create a verifiable trail regarding its origin. Advancements in watermarking technology ensure that watermarks remain detectable even after common modifications, such as cropping, applying filters, altering frame rates, and saving with various compression schemes.[42] While it provides a valuable line of defense to uphold content integrity and is gaining traction among policymakers and governments, *it is not a silver bullet*.[43] Sophisticated adversaries can remove or alter watermarks without significantly degrading the quality of media.[44] In the case of deepfake audios, "the acoustic and phone channel degradations make watermarking more vulnerable to attacks," a finding that aligns with a study on the ability of bad actors to "wash out" the watermarking.[45,46] While watermarking all AI-generated content might be unrealistic, pursuing this outcome is nonetheless worthwhile as a partial measure.

Content provenance is a growing field that has also gained significant traction. This approach involves hard-coding metadata into media at the time of creation, which provides an alternative method for verifying the authenticity of digital content. By embedding tamper-proof information directly within the media file, it is possible to trace its origin and verify its integrity. Such metadata includes details about the content's creation, edits and distribution, enabling a more robust validation process. Several initiatives, such as those spearheaded by C2PA,

42 "SynthID: Identifying AI-generated content with SynthID," Google DeepMind, accessed September 13, 2024, https://deepmind.google/technologies/synthid/.

43 Kevin Klyman and Renée DiResta, "Beyond Watermarks: Content Integrity Through Tiered Defense," Council on Foreign Relations, May 8, 2024, https://www.cfr.org/blog/beyond-watermarks-content-integrity-through-tiered-defense.

44 Nick Gaubitch, "Does Watermarking Protect Against Deepfake Attacks?" *Pindrop*, October 20, 2023, https://www.pindrop.com/blog/does-watermarking-protect-against-deepfake-attacks.

45 Gaubitch, "Does Watermarking Protect Against Deepfake Attacks?"

46 Kate Knibbs, "Researchers Tested AI Watermarks—and Broke All of Them," *Wired*, October 2, 2023, https://www.wired.com/story/artificial-intelligence-watermarking-issues/.

Truepic, and Starling Lab, are at the forefront of developing frameworks and open standards for identifying the authenticity of digital content.[47,48,49]

While not accessible to individual users, there are other detection technologies and methods that can be implemented at the content platform-level that are increasingly adept at identifying deepfakes, including convolutional neural networks (CNNs) and diffusion models.[50] Platforms like Microsoft Azure Sentinel and Deepware have also made improvements in detection accuracy to flag potentially deceptive content via CNN architecture.[51] While the widespread adoption of these approaches and technologies is still in its nascent stages, their development represents a promising advancement in digital content verification and complements existing detection strategies.

→ **Recommendation C**: Modernize authentication approaches to account for AI.

Authentication has long relied on "something you know" (e.g., a password, phrase, PIN, or answers to security questions), and more recently augmented with "something you have" (e.g., a one-time code from an authentication app or sent via text message, or cryptographic identification card/device/token), and possibly "something you are" (e.g., biometric, such as fingerprint or facial recognition). While passwords have long been in dubious standing due to their re-use, countless data breaches, and ease of brute forcing, "something you know" is now even less reliable due to AI's ability to summarize and recall information from large datasets, to include scrapes of public social media, public records, and pooled data breach takings. Users should assume that answers to their security questions, typically used for password resets and account recovery, can be known by bad actors using AI capabilities. And now generative AI has been demonstrated to fool some biometric authentication methods, putting the reliability of "something you are" at risk as well.

That leaves "something you have." Last year, industry veterans argued for broader adoption of public key cryptography as a deterministic factor in authentication to counter probabilistic factors like biometrics that can be vulnerable to AI-powered attacks.[52] This is the technology already employed by U.S. government employees when using a Personal Identity Verification

47 "Coalition for Content Provenance and Authenticity," C2PA, accessed August 25, 2024, https://c2pa.org/.
48 "Authenticity infrastructure for the internet," Truepic, accessed August 25, 2024, https://truepic.com/.
49 "Introducing: The Starling Framework for Data Integrity," Starling Lab, accessed August 25, 2024, https://www.starlinglab.org/.
50 Negar Kamali et al., "How to Distinguish AI-Generated Images from Authentic Photographs," Northwestern University, June 2024, https://arxiv.org/abs/2406.08651.
51 Franklin Okeke, "7 Best AI Deepfake Detector Tools For 2024," Techopedia, January 23, 2024, https://www.techopedia.com/best-ai-deepfake-detectors.
52 Jeremy Grant and Zack Martin, "AI is here: What it means for digital identity and cybersecurity," Center for Cybersecurity Policy and Law, August 2, 2023, https://www.centerforcybersecuritypolicy.org/insights-and-research/ai-is-here-what-it-means-for-digital-identity-and-cybersecurity.

(PIV) card or Common Access Card (CAC) when accessing a government facility or logging onto a computer system, both of which use public key infrastructure (PKI) certificates for secure login, and an essential element of phishing-resistant multi-factor authentication called for in the U.S. government's zero trust architecture strategy.[53] Beyond PKI, the Fast Identity Online (FIDO) Alliance has developed a more lightweight approach to authentication leveraging public key cryptography called FIDO2, which is supported by every major technology platform and accommodates both passwordless authentication and traditional MFA via security keys such as YubiKeys. Organizations outside of government should adopt these physical authentication solutions.

Apart from hardening authentication, AI is also threatening the tools we use online for identity verification—the process people go through when they first establish an account, or look to recover it after losing a password or other authenticator. At the individual level, in-person identity verification within the United States typically involves a state-level motor vehicle office, which issues driver's licenses and state identification cards, or the U.S. Department of State, which issues U.S. passports. Both require an applicant to be physically present and to provide identity documents, such as a birth certificate or social security card.

In an effort to leverage the robust in-person experience to address online challenges, numerous states have now started to implement mobile driver licenses (mDLs), a PKI-based digital version of their physical credential that resides on a mobile device. mDLs are initially being used only in high-trust authentication and authorization use cases for in-person transactions; for instance, they can be used to clear through airport security to board a flight. But a new NIST initiative at the National Cybersecurity Center of Excellence (NCCoE) is developing standards and best practices to extend the power of mDLs to support online use cases that require robust, remote identity proofing, such as in financial services, government services, and health care.[54] This will enable organizations to extend the benefits of public key cryptography beyond authentication to also support more robust online identity verification use cases.

Finally, one must consider how these approaches can hold up in contingencies, such as an individual losing their smartphone or YubiKey or having their phone and wallet stolen while on foreign travel. Our identity ecosystem must be able to account for these situations and have a human-based recovery protocol to re-establish the basis of trust. This might involve the individual going to a local motor vehicle office, post office, embassy, or consulate to be re-verified, leading to a secure recovery code being conveyed directly to the service provider

53 Shalanda D. Young, "Memorandum for the Heads of Executive Departments and Agencies: Moving the U.S. Government Toward Zero Trust Cybersecurity Principles,"White House Office of Management and Budget, M-22-09, January 26, 2022, https://www.whitehouse.gov/wp-content/uploads/2022/01/M-22-09.pdf.

54 "Digital Identities - Mobile Driver's License (mDL)," National Cybersecurity Center of Excellence, National Institute of Standards and Technology (NIST), accessed August 18, 2024, https://www.nccoe.nist.gov/projects/digital-identities-mdl.

who can then restore the user's access. This report encourages further work by relevant U.S. departments and agencies, state officials, certificate authorities, industry coalitions, and standards bodies to develop and implement needed solutions.

Recommendation D: Educate society to navigate the challenges brought by AI deepfakes.

Today, it is all too easy to mistake AI-generated deceptive audio or images for authentic content, as sophisticated impersonation techniques increasingly obscure reality. Strengthening our defenses against social engineering requires a concerted effort to equip individuals and organizations with the necessary skills, knowledge, and tools to recognize AI-driven deception.

Cybersecurity awareness and digital literacy are the first factors in this equation: educating users across all walks of life about the deceptive capabilities unlocked by AI models, and how to protect against them. As social engineering is a socio-technical problem and deeply rooted in human behavior, deepfake awareness campaigns should encourage a culture of "mindful content consumption," including critical thinking.[55] Some of the available tools that would support awareness include Google's "About this image" feature, which helps users proactively check content they are consuming online.[56] However, awareness alone is not sufficient, since it does not make the detection of AI-generated content easier. Even for the trained eye, it is becoming increasingly difficult to validate AI-generated outputs.

This brings us to the second factor of the security equation: investing in advanced deepfake detection tools and capabilities at the platform level to flag inauthentic content for the user. Platforms should utilize adaptive detection tools that can flag and alert users of AI-generated deceptive content, instead of amplifying them for clickbait through recommendation algorithms.

Premise #3: AI will make software vastly more secure.

Organizations have made significant strides in advancing software security through frameworks such as NIST's Secure Software Development Framework and various

55 "Sifting Through the Pandemic: Information Hygiene for the Covid-19 Infodemic," *Infodemic Blog*, accessed September 18, 2024, https://infodemic.blog/.

56 Nidhi Hebar and Christopher Savcak, "3 new ways to check images and sources online," Google, October 25, 2023, https://blog.google/products/search/google-search-new-fact-checking-features/.

"secure-by-design" initiatives.[57,58] Nonetheless, software security remains a daunting challenge due to the immense volume and complexity of code within modern software. For example, Microsoft Windows 11 contains around 50 million lines of code, far more than a single human can review over an entire career. Human error and the difficulty of spotting vulnerabilities further exacerbate the challenges of getting software security right. Since models have improved the speed at which code can be generated, it should therefore be no surprise that leveraging AI to write, review, and improve code has quickly become a popular cause within industry and beyond, leading many to hope that AI might alleviate all manners of software-related challenges. Is this an unrealistic panacea, or is there something to it?

One does not have to look toward the distant future to see the beginnings of this promising vision. A recent report by the Aspen Institute outlined an optimistic future scenario in which AI greatly improves cybersecurity.[59] As part of this scenario, the report characterized six relevant use cases where AI could improve the security and quality of both existing and new code: code assessment, code recommendations, code monitoring, forecasting, rewriting code, and code automation.[60]

USING AI TO FIND AND FIX VULNERABILITIES IN SOURCE CODE

Indeed, industry incumbents and startups alike are racing to develop AI coding capabilities, making surprisingly rapid progress. In fact, several startups have started experimenting with autonomous agents capable of sifting through "open-source code bases, find vulnerabilities and fix them."[61] During a recent Defense Advanced Research Projects Agency (DARPA) competition, participants found 22 unique vulnerabilities across widely-used open-source software components, 15 of which were automatically patched, and one of which was a zero-day. According to Dan Guido, Founder of Trails of Bits, "there [is] just too much code to look through, and it [is] too complex to process in order to find all the vulnerabilities that are spread out", but "AI is an opportunity that might help assist us in finding and fixing security issues that are now pervasive and increasing in number."[62]

57 "Secure Software Development Framework," Computer Security Resource Center, NIST, February 25, 2021, https://csrc.nist.gov/Projects/ssdf.

58 "Secure by Design," Cybersecurity and Infrastructure Security Agency (CISA), accessed August 16, 2024, https://www.cisa.gov/securebydesign.

59 "Envisioning Cyber Futures with A.I.," Aspen Institute.

60 "Envisioning Cyber Futures with A.I.," Aspen Institute.

61 Christian Vasquez, "DARPA competition shows promise of using AI to find and patch bugs," *CyberScoop*, August 12, 2024, https://cyberscoop.com/darpa-competition-shows-promise-of-using-ai-to-find-and-patch-bugs/.

62 Vasquez, "DARPA competition."

USING AI TO CONDUCT FUZZ TESTING

Fuzz testing, or fuzzing, is another technique to discover security vulnerabilities or bugs in the source code of software applications by introducing random inputs in an effort to crash it, and thus identify faults that would otherwise not be apparent.[63] In a blog post dated nearly five years ago, cybersecurity company Fortinet asserted that "[a]pplying AI and machine learning models to fuzzing will enable it to become more efficient and effective" and that "[b]lack hat criminals will be able to develop and train fuzzing programs to automate and accelerate the discovery of Zero-Day attacks."[64]

This prediction was on target, as Google's open source security team more recently wrote that its "OSS-Fuzz" automated vulnerability discovery tool for open source projects, in use since 2016, might be fully automated by leveraging the company's LLMs to write fuzz targets—functions that use randomized input to test the targeted code. The experiment generated promising results with iterative increases in the AI-enhanced fuzzing's code coverage increasing to promising levels.[65] This is particularly useful for neglected software, such as certain open source libraries that are commonly incorporated into other products.[66]

USING AI TO ASSIST IN WRITING CODE

AI coding assistants, which make use of large language models (LLMs) and natural language processing to assist developers through the coding process, are becoming increasingly common. Github's Copilot, Google's Gemini Code Assist, and IBM's watsonx Code Assistant are just a few examples that highlight this rapidly growing product segment, which is poised to transform the future of software development. With openly accessible code repositories (i.e., Hugging Face, GitHub) as well as an overreliance on a handful of software companies—which leads to risk concentration—it is increasingly difficult to answer the following question: who is responsible for the maintenance of secure software?

Such AI code assistants have already demonstrated their ability to streamline workflows, enhance code readability, and automate routine tasks, resulting in notable improvements in overall productivity. In conversation with IST, one industry heavyweight identified that the two most transformative applications of current AI tooling are its role in helping new personnel

63 "What is fuzz testing?" GitLab, accessed August 24, 2024, https://about.gitlab.com/topics/devsecops/what-is-fuzz-testing/.

64 FortiGuard SE Team, "Predictions: AI Fuzzing and Machine Learning Poisoning," Fortinet, November 15 2018, https://www.fortinet.com/blog/industry-trends/predictions--ai-fuzzing-and-machine-learning-poisoning-.

65 Dongge Liu, Jonathan Metzman, and Oliver Chang, "AI-Powered Fuzzing: Breaking the Bug Hunting Barrier," Google Security Blog, August 2023, https://security.googleblog.com/2023/08/ai-powered-fuzzing-breaking-bug-hunting.html.

66 Zoë Brammer et al., "Castles Built on Sand: Towards Securing the Open-Source Software Ecosystem," Institute for Security and Technology, April 2023, https://securityandtechnology.org/virtual-library/reports/castles-built-on-sand-towards-securing-the-open-source-software-ecosystem/.

comprehend legacy code, and the production of boilerplate code.[67] In another application, researchers from the Research, Development and Innovation Centre at the University of Campina Grande found that AI models like ChatGPT are increasingly agile at "pinpointing syntax and semantic issues [within source codes] while proposing potential solutions" to ameliorate them.[68] These AI models also facilitate more efficient code reviews, saving time and boosting software quality.[69]

IST's survey confirmed that some organizations are actively deploying AI tools to assist in writing and reviewing code while identifying potential security flaws, with one respondent from a global cybersecurity company noting that their pilot started as early as six months ago. Another respondent from a leading technology company reported that AI tools have significantly enhanced and supplemented their code review and writing processes. This integration of AI has not only mitigated the impact of skills shortages, but also optimized operational efficiency and accuracy. The respondent emphasized that "these skills [have been made] more obtainable via generative AI even despite its imperfections [which] is incredibly valuable [...] these are major, major improvements."

Finally, another compelling use case is leveraging LLMs to rewrite legacy code from languages with known memory safety issues, such as C and C++, into memory-safe languages like Rust. However, this key cybersecurity imperative brings a heavy effort and cost, as converting large codebases—like an operating system—might take hundreds of programmers working full time for years. The DARPA's Translating All C to Rust (TRACTOR) program aims to accelerate these efforts by automating the translation of legacy C code to Rust with the help of LLM-powered solutions.[70] Additionally, researchers recently published the results of their test of the ability of "five state-of-the-art LLMs to translate C and Go code taken from real-world projects to Rust," concluding that the "LLMs are indeed capable of translating code to Rust, though there [is] room for improvement."[71] With continued work and investment, this capability has the potential to significantly accelerate the retirement of unsafe software languages, yielding ecosystem-wide security benefits.

---

67 Boilerplate code refers to sections of code that are repetitively used in various parts of a program with little to no modification. It often includes standard, routine code that can be repurposed for different situations. "What is Boilerplate Code?," Amazon Web Services, accessed August 20, 2024, https://aws.amazon.com/what-is/boilerplate-code/.

68 Yonatha Almeida et al., "AICodeReview: Advancing Code Quality with AI-Enhanced Reviews," *SoftwareX*, May 2024, https://www.sciencedirect.com/science/article/pii/S2352711024000487.

69 Yonatha Almeida et al., "AICodeReview."

70 Clarence Oxford, "Eliminating Memory Safety Vulnerabilities with Rust and AI," *Space Daily*, August 2, 2024, https://www.spacedaily.com/reports/Eliminating_Memory_Safety_Vulnerabilities_with_Rust_and_AI_999.html.

71 Hassan Ferit Eniser et al., "Towards Translating Real-World Code with LLMs: A Study of Translating to Rust," arXiv, accessed October 1, 2024, https://arxiv.org/html/2405.11514v2.

HOWEVER, ON THE DOWNSIDE...

While AI could enhance software security through code assistance, vulnerability management, and fuzzing, these very capabilities could be harnessed by bad actors for malicious ends.

Additionally, researchers studying the effects of AI code assistants in controlled environments have raised concerns related to human factors, particularly the tendency to over-rely on tools that enhance efficiency. Some have highlighted that current evaluations of AI models often emphasize code correctness while overlooking its security implications, conflating the two distinct, yet crucial objectives.[72] As organizations integrate AI code assistants into their processes, it is essential to avoid equating the efficiency gains provided by AI with genuine improvements in code accuracy and security.

AI code assistants have also, in some cases, introduced security risks and generated insecure code in controlled environments.[73] Given the size and complexity of modern software packages, some coding errors are inevitable, potentially creating vulnerabilities that could be exploited by malicious actors in practice. For instance, Stanford University researchers found that "participants with access to an AI assistant wrote insecure solutions more often than those without access to an AI assistant for four of [...] five programming tasks."[74] This suggests that while AI assistants can lower barriers for developers and boost efficiency, they might also lead to a "false sense of security."[75]

Furthermore, a study conducted by researchers at the University of California San Diego highlights yet another challenge: participants found it more difficult to identify errors and debug code generated by Copilot compared to code they wrote themselves.[76] Unfamiliarity with AI-generated code can impede effective debugging and error detection if developers lack the context and understanding they have with their own code. Given these examples, concerns remain about the practical application of AI code assistants and the risk of increased oversight lapses or the replication of security vulnerabilities. We should not conflate AI's efficiency in generating code with its ability to generate secure and accurate code; achieving both is essential.

In industry settings, one IST survey respondent cautioned against using AI code assistants for applications in high-trust contexts. These include integrating AI-generated code into

72 Mohammed Latif et al, "Generate and Pray: Using SALLMS to Evaluate the Security of LLM Generated Code," arXiv, June 3, 2024, https://ui.adsabs.harvard.edu/abs/2023arXiv231100889S/abstract.

73 Neil Perry et al, "Do Users Write More Insecure Code with AI Assistants?" arXiv, December 18, 2023, https://arxiv.org/pdf/2211.03622.

74 Perry et al., "Do Users Write More Insecure Code with AI Assistants?"

75 Perry et al., "Do Users Write More Insecure Code with AI Assistants?"

76 Shraddha Barke, Michael B. James, and Nadia Polikarpova, "Grounded Copilot: How Programmers Interact with Code-Generating Models," *Proceedings of the ACM on Programming Languages* 1, article 1 (April 2018), https://arxiv.org/pdf/2206.15000.

components like kernel modules or firewalls, where the consequences of failure can be severe. While organizations are understandably eager to adopt AI applications, this sentiment highlights the importance of intentional and risk-informed deployment. For example, utilizing LLMs for tasks with wider margins for error can help mitigate these risks. One example is fuzz testing: if the model fails, the only consequence is the failure of the test itself, not the compromise of critical system components. This approach allows organizations to experiment with lower-risk AI use cases and incrementally increase task complexity and risk tolerance over time. Moreover, the respondent recommended there be ways to audit the origins of code, particularly code generated by LLMs. This is a particularly compelling finding, as auditing—whether for open-source libraries, specific human programmers, or LLM copilots—is becoming increasingly relevant in discussions around transparency and accountability.

It may be that AI's ability to fix and improve code safety at speed and scale far outweighs the potential of it inadvertently introducing new bugs. This is an area ripe for continued research.

Recommendation E: Optimize both human and AI resources to achieve efficiency and software quality.

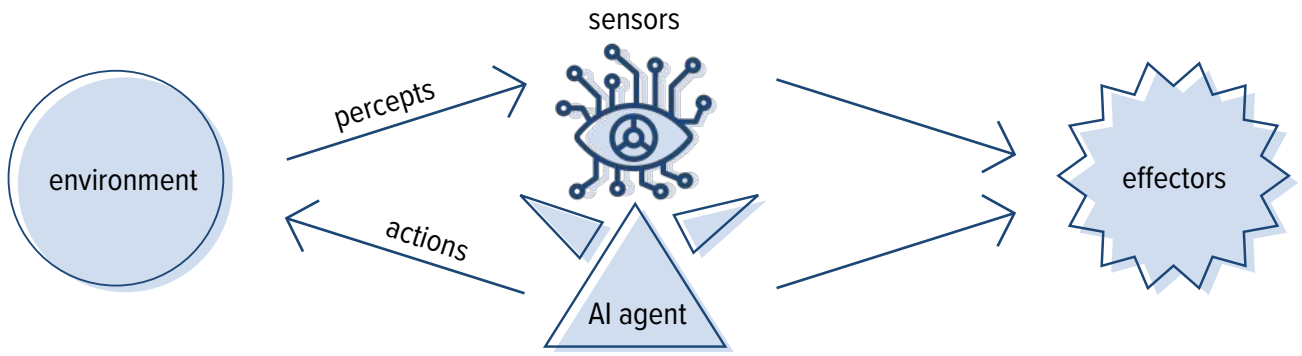
AI code assistants are already proving to be a powerful tool in the software development process, acting as a force multiplier by amplifying human productivity to new levels. While the integration of AI tools into code generation would enhance productivity and efficiency, it introduces the risk of over-reliance on these outputs. Our ability to critically assess AI-generated code is highly dependent on ensuring we still acquire the necessary skills to know what went wrong or how to write the code more securely.

From this perspective, striking the right balance between automating the code generation process and vetting its correctness and robustness through human expertise will determine AI's potential to enhance software security. By integrating AI assistants in a way that complements human judgment, organizations can achieve a synergy where efficiency enhances quality, accuracy, and security, rather than compromising them. In one such case, a survey respondent from a global cybersecurity company remarked on how they have prioritized quality assurance at the center of their deployment strategy. By mandating feedback from staff who are working with AI tools and asking targeted questions—"What worked and didn't work? What was helpful and unhelpful?"—they have implemented a positive feedback loop that accounts for the limitations and caveats associated with these tools.

The Next Leap: Agentic AI in Cybersecurity

Agentic AI, or AI agents, refers to advanced software systems that can analyze information, make decisions, and plan actions autonomously.[77] These systems leverage intricate algorithms to assess various options and select optimal paths to act on their own. Equipped with sensors—ranging from physical devices like cameras to virtual tools such as data access—these agents can perceive their environments. Their "effectors"[78] enable them to act, whether through physical means or by sending commands to other software.[79]

In cybersecurity, agentic AI could play a crucial role by continuously adapting to the evolving landscape of threats. With varying levels of autonomy, these agents can detect anomalies, respond to breaches, and implement protective measures, all while requiring minimal human intervention in lower-risk environments. While there is widespread enthusiasm for advancements in AI agents' capabilities, there remain several limitations to the accuracy of their outputs. (See Our Watch List)



Premise #4: AI will revolutionize security operations.

System owners *should* have a comparative advantage over attackers in one key aspect: they understand the internal terrain, data stored within the network, and what "normal" network traffic looks like. But this potential advantage is often not realized, as many defenders are drowning in data, alerts, and network administration tasks that flood their systems.

ML has long been integrated into cybersecurity operations, including within Security Information and Event Management (SIEM) and Security Orchestration, Automation, and Response (SOAR) solutions. Equally as impactful is AI's role as a force multiplier in

77 Lareina Yee, Michael Chui, and Roger Roberts, "Why agents are the next frontier of generative AI," *McKinsey Quarterly,* July 24, 2024, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/why-agents-are-the-next-frontier-of-generative-ai.

78 Effectors are any devices that affect the physical environment. An actuator is the actual mechanism that enables effectors to execute an action.

79 "What are AI Agents?," Amazon Web Services, accessed September 18, 2024, https://aws.amazon.com/what-is/ai-agents/.

strengthening the cyber workforce, enabling security professionals to leverage advanced tools, streamline complex tasks, and focus on higher-level strategic decision making rather than routine operations. What does this mean for the day-to-day defender?

AI can help turn the tables by easily managing big-data and gleaning insights that can keep analysts focused on complex tasks while offloading the operational minutiae. By automating routine operations—such as network traffic analysis, file classification, and endpoint protection—AI tools alleviate the burden on analysts, enabling them to tackle more nuanced queries with greater precision. AI "case assistants," for example, have made it easier for analysts to identify embedded malicious code. Across the board, LLMs have simplified data comprehension for all security personnel, though their impact has been particularly transformative for Tier 1 Security Operations Center (SOC) analysts during case investigations by reducing the time needed to accurately identify malicious intent.

"It's going to be an arms race, but defenders still have an advantage."

-IST survey response

However, the implementation of AI tools is not without challenges. According to a respondent from a global cybersecurity company reflecting on outputs from their recently implemented AI case assistant, "some summaries are fantastic, some good, some are junk," but certain use cases "really have legs." One of these promising applications is making data accessible through a natural language query, instead of an analyst having to author complex SQL queries (a skill some reportedly lack). At present, the company's use of LLMs within the SOC mostly involves "stitching together more deterministic actions, routing between functions, but not letting the LLM choose its own adventure" or carrying out automated responses. Consistent with sentiments reflected by others surveyed, the respondent sees a potential to automate many tier 1 SOC functions, freeing up human resources to focus on higher-order (i.e., tier 2 and beyond) tasks, and so on up the stack to increase security value. While the idea of an automated "SOC analyst in a box" is still a few years away, it is clear that as AI capabilities advance, they will affect operational efficiencies as well as the hierarchies within SOCs, emphasizing the importance of human oversight in the interim.

In summary, AI tooling enhances SOC efficiency and effectiveness. By addressing and eliminating common bottlenecks such as those encountered during triage and investigations, AI's ability to quickly sift through and analyze vast amounts of data transforms data from a traditional burden to a strategic asset, allowing defenders to diagnose and respond to threats with increased speed and accuracy.

Another promising use case involves AI conducting an automated network inventory and network auditing, searching for unpatched or misconfigured systems, detecting

unauthorized devices (aka, "shadow IT"), and the like. As noted by a survey respondent from a large cybersecurity company, AI tooling can "enable [a] higher fidelity understanding of your environment and allow you to process and act on your information more quickly." The respondent also noted that the integration and use of AI models to enhance security operations will indeed improve the offense-defense balance: "It's going to be an arms race, but defenders still have an advantage."

Often, defenders may not fully leverage this advantage—a lack of incentives, resources, or a blind eye toward sustainable security practices can undermine the assets that defenders could otherwise capitalize on, leading to missed opportunities to leverage the data advantage that system owners inherently possess. As one respondent from a global cybersecurity company noted, given the advent of AI tools, defenders stand to gain "a fuller understanding of their environment than any attacker will."

→ **Recommendation F**: Integrate AI into security operations workflows, but protect your models.

While AI models hold the promise of enhancing defenders' understanding and security of their networks, access to AI alone is not enough to drive a security revolution. Two critical factors will determine the success and amplitude of this revolution: the effective *integration* of AI models into defenders' processes, as well as their ability to *maintain* the security of these models against adversarial attacks. Researchers have found that AI models could be subject to more than 30 different attack vectors that adversaries could adopt, and the proper protection of these models is urgent.[80]

- **Integrating AI models**. The potential of AI in cyber defense is highly contingent on how well AI models are integrated into existing organizational frameworks and processes. Effective integration means not only incorporating AI tools into security operations, but ensuring that these tools enhance system security while preserving essential human judgment, and that they are well intertwined with the organization's infrastructure and networks. In other words, organizations must achieve synergy between traditional security measures and the advanced capabilities of AI, optimizing for speed, scale, and efficiency without sacrificing the nuanced decision-making that human oversight provides.
- **Protecting AI models**. AI's role in cyber defense is not limited to its integration within security systems—it also requires ongoing protection from adversarial threats such as model and data poisoning, prompt injection, membership inference attacks, and

80 Sella Nevo et al., "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models," RAND, May 30 2024, https://www.rand.org/pubs/research_reports/RRA2849-1.html.

jailbreaking. Addressing these risks demands a comprehensive approach to securing the AI lifecycle, an area where initiatives like Google's Secure AI Framework (SAIF) have been particularly compelling. SAIF, for example, emphasizes the need for robust data governance and lifecycle management to safeguard AI training datasets, ensuring that the data feeding these models remains accurate, reliable, and resistant to tampering or potential misuse, further underscoring the need for ongoing research to track emerging risks against AI models and develop effective countermeasures.[81] For example, practices such as red teaming—the simulation of attacks on AI systems to identify vulnerabilities and assess the effectiveness of defensive guardrails—have become increasingly common for gaining insights into model security. Beyond just securing new, state-of-the-art AI applications and tools, survey participants noted that their client services increasingly involve leading smaller organizations into the AI space more responsibly, with a focus on governance and architectural hardening. As adversarial techniques evolve, so too must the defensive measures and protocols designed to protect AI systems.

> *"Where there is a lack of wisdom or experience, AI could be a suit of armor."*
>
> - IST survey response

Ultimately, effective cyber defense will demand more than just technological solutions—it requires ongoing efforts in personnel training and robust model management to combat the threats discussed throughout this report. As AI continues to be adopted across the industry, maintaining the security of AI must remain an adaptive process, ensuring that both AI models and the people managing them are equipped to face an evolving threat landscape.[82]

81 "Google's Secure AI Framework (SAIF)," Google, accessed August 2, 2024, https://safety.google/cybersecurity-advancements/saif/.

82 "Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector," U.S. Department of the Treasury, March 2024, https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf.

The Great Equalizer: Empowering the Cybersecurity Workforce

AI is addressing long-standing cyber workforce challenges such as skill shortages, talent gaps, and time constraints.[83] AI tools, both upstream and downstream, will trigger a long-term, force multiplier effect by equipping security professionals with greater access to information and alleviating traditional burdens.[84] These tools streamline workflows, foster greater cross-team collaboration, and automate menial tasks, allowing security professionals to focus on more strategic and complex issues.

IST's survey data corroborates this enthusiasm: large cybersecurity providers are emphasizing the role of AI in delivering more efficient, cost-effective, and customized solutions for their customers. Overall, AI integration has seen an upturn across organizations of all sizes. Investments in training programs are also on the rise, aimed at equipping human representatives, including non-technical staff, with the skills needed to effectively utilize AI. Notably, small and medium-sized businesses (SMBs), often limited by resources and expertise, stand to gain significantly from AI's democratizing effect. AI tools can help SMBs better manage cyber-risks by lowering the skill barrier, enabling non-technical personnel to make informed decisions by translating big data into actionable insights. As one industry respondent aptly put it, "where there is a lack of wisdom or experience, AI could be a suit of armor."

Premise #5: AI is supercharging adversarial reconnaissance and target identification.

AI is transforming the speed and scale of adversarial reconnaissance, enabling threat actors to rapidly identify targets, gather intelligence, and leverage this data for future operations. With AI tools, adversaries can sift through vast amounts of data, automate the collection of insights, and extract detailed information about those they wish to target—all without human intervention. In doing so, AI-enhanced reconnaissance improves adversaries' ability to prioritize targets and sharpens their understanding of attack surfaces and how to penetrate defenses.

The first section examines these organizational reconnaissance strategies, rooted in documented, real-world applications of how threat actors have leveraged LLMs in this context.

83 "2023 Global Future of Cyber Survey," Deloitte Global, October 2023, https://www.deloitte.com/global/en/services/risk-advisory/content/future-of-cyber.html.

84 Mike Morros, Arun Perinkolam, and Jacob Crisp, "Entering the Era of Generative AI-Enabled Security," Google Cloud and Deloitte, 2023, https://services.google.com/fh/files/misc/entering-the-era-of-gen-ai-enabled-security.pdf.

The second section, technical reconnaissance—though less substantiated at present—will likely become an emerging challenge as AI model capabilities improve.

USING AI TO CONDUCT ORGANIZATIONAL RECONNAISSANCE

According to the United Kingdom's National Cyber Security Centre, AI-driven reconnaissance is becoming more invasive and effective, as bad actors may use AI models to analyze vast amounts of data with increased speed and scale.[85] A surveyed organization highlighted that AI tools can enable threat actors to observe their targets meticulously and identify vulnerabilities with increased precision, allowing them to refine and enhance their cyber-attack strategies. Other survey respondents corroborated this observation, noting that AI tools can facilitate thorough research, enabling bad actors to meticulously observe and compare potential targets, ultimately identifying vulnerabilities with increased precision. This marks a significant leap in efficiency compared to traditional reconnaissance methods, as AI shifts the focus from broad, indiscriminate probing to highly targeted exploitation.[86]

As bad actors continue to experiment with AI tools, researchers have identified additional use cases that enhance reconnaissance capabilities.[87] Two notable advancements are intelligent profiling and AI-driven data collection. Supported by AI technologies like fuzzy logic, machine learning, and convolutional neural networks, intelligent profiling enables attackers to analyze social media and other data sources with extreme precision, refining their ability to target specific individuals or organizations more effectively. This significantly improves message targeting, profiling, and personalization.[88,89] In parallel, AI-driven data collection, powered by natural language processing and deep neural networks (DNNs), automates the gathering of general and specific information relevant to cyber threats.

Bad actors are already using LLMs to gain detailed insights into potential targets, allowing them to prioritize attacks and victim selection based on a comparative analysis of vulnerabilities.[90] State-affiliated malicious actors, in particular, are increasingly leveraging AI tools in cyberspace for reconnaissance and other phases of their cyber operations.

---

85 "The Near-Term Impact of AI on the Cyber Threat," National Cyber Security Centre UK, January 2023, https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat#section_1.

86 "The Near-Term Impact of AI on the Cyber Threat," National Cyber Security Centre UK.

87 Guembe Blessing, Ambrose Azeta, and Sanjay Misra, "The Emerging Threat of AI-Driven Cyber Attacks: A Review," *Applied Artifical Intelligence* 26 (2022), January 28, 2022, https://www.researchgate.net/publication/359038562_The_Emerging_Threat_of_Ai-driven_Cyber_Attacks_A_Review.

88 Muhammad Bilal et al., "Social Profiling: A Review, Taxonomy, and Challenges," *Cyberpsychology, Behavior, and Social Networking* 22, no. 7 (July 2019): 433–50, https://doi.org/10.1089/cyber.2018.0670.

89 Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," arXiv, February 20, 2018, https://doi.org/10.17863/cam.22520.

90 "Threat Report: How Threat Actors Are Leveraging Artificial Intelligence (AI) Technology to Conduct Sophisticated Attacks," Global Threat Intelligence by Deloitte, 2024, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-design-ai-threat-report-v2.pdf.

A collaborative study by OpenAI and Microsoft in February 2024 highlighted how AI, including tools like ChatGPT, has been deployed by threat actors from Iran, the Democratic People's Republic of Korea (DPRK), Russia, and the People's Republic of China (PRC).[91] Microsoft tracked threat actor groups such as Forest Blizzard (Russia), Emerald Sleet (DPRK), Salmon Typhoon (PRC), and Charcoal Typhoon (PRC) who have been observed deploying AI to enhance their reconnaissance capabilities.[92] Threat actors' LLM-informed queries sought actionable insights on a variety of technologies and potential vulnerabilities, such as radar imaging technologies, satellite capabilities, and even domestic issues within targeted countries.[93]

For instance, Charcoal Typhoon, a China-nexus threat group, has reportedly been using ChatGPT to learn more about different companies and cybersecurity tools "indicative of preliminary information-gathering stages."[94] Another China-nexus group, Salmon Typhoon, has leveraged OpenAI to search for open-source information about a variety of intelligence organizations and other regional actors. In the Asia-Pacific region, Emerald Sleet has been observed utilizing the model to identify defense-related organizations and explore publicly-available vulnerability repositories.

USING AI TO CONDUCT TECHNICAL RECONNAISSANCE

Bad actors' reconnaissance and probing efforts will be increasingly automated with AI capabilities that can provide near ubiquitous and near real-time coverage of every device exposed to the public internet. In light of this new reality, we are left with several unsettling questions: Are our networks ever safe, even when we abide by security best-practices? Are we even able to detect or contain AI-enabled probing efforts before it is too late? Or are we merely waiting for the next compromise to adapt our perimeter security?

AI-enabled reconnaissance—which is always active, and does not tire nor require real-time human supervision—requires a full reckoning of an organization's external attack surface. This constant and tireless surveillance means that there is no way to hide vulnerabilities in external-facing assets, necessitating that organizations match this reconnaissance with an optimized and well-maintained, proactive defense.

Attackers are leveraging AI to identify gaps across organizations' ever-expanding attack surfaces. Active reconnaissance, a systematic process that involves probing systems and

91 "Cyber Threat Awareness Report: Lockbit Takedown Impact & AI's Dual Role in Cybersecurity," CVP, March 4, 2024, https://www.cvpcorp.com/cyber-blog/cyber-threat-awareness-report-march-04-2024.

92 "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors," OpenAI, February 14, 2024, https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/.

93 "Staying ahead of threat actors in the age of AI," Microsoft, February 14, 2024, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

94 "Staying ahead of threat actors in the age of AI," Microsoft.

networks via different techniques like port scanning, vulnerability scanning, and enumeration, can guide adversaries in selecting attack vectors and strategies for further exploitation.[95]

From on-premises infrastructure to the cloud, managing this expansive attack surface has become a mounting challenge given the explosive growth of organizations' assets. The sheer number of physical systems, cloud infrastructure, Application Programming Interfaces (APIs) Software as a Service (SaaS) applications, and Internet of Things (IoT) devices has exponentially expanded the attack surface, and when left unmonitored, unpatched, or simply forgotten about, can become new opportunities for threat actors to exploit. Indeed, many companies may lack full visibility into how many assets they have, and without sufficient security practices, may grant adversaries the upperhand.[96]

"Are our networks ever safe, even when we abide by security best-practices? Are we even able to detect or contain AI-enabled probing efforts before it is too late? Or are we merely waiting for the next compromise to adapt our perimeter security?"

- The Implications of Artificial Intelligence in Cybersecurity

Threat actors have already been observed taking advantage of AI to surveil and scan the various assets mentioned. As a recent example, threat actors were reported exploiting vulnerabilities in the ServiceNow platform by "using automated tools to target login pages, aiming to deploy with two payloads. The first to test if remote code execution is possible and the second to reveal database users and their passwords."[97] While this does not represent a novel capability or outcome—insofar as a human can also conduct such reconnaissance manually—it does represent a significant leap in speed, scale, and completeness.

95 "What is Cyber Threat Intelligence?" SentinelOne, November 16, 2023, https://www.sentinelone.com/cybersecurity-101/threat-intelligence/what-is-cyber-reconnaissance/.

96 Franklin Okeke, "Without Knowing Your APIs, You Don't Know Your Attack Surface: Akamai Interview," Techopedia, June 13, 2024, https://www.techopedia.com/api-security-depends-on-api-visibility-akamai-interview.

97 Solomon Klappholz, "Critical ServiceNow vulnerabilities exploited in 'global reconnaissance campaign,'" *IT Pro*, August 2, 2024, https://www.itpro.com/security/critical-servicenow-vulnerabilities-exploited-in-global-reconnaissance-campaign.

AI-Powered Vulnerability Mapping: A Tactical Tool for Threat Actors and Defenders Alike

AI's impact on vulnerability mapping could be particularly striking. For threat actors, AI can uncover patterns of vulnerabilities within target networks, enabling them to identify and exploit weaknesses more efficiently. Conversely, the same capability can empower organizations to proactively address and detect these vulnerabilities before they are exploited.

No longer confined to traditional methods, bad actors could potentially leverage AI models to accelerate the identification of vulnerabilities and detect patterns that were previously hidden. By feeding AI models with vast repositories of known software vulnerabilities and exploits, adversaries can efficiently pinpoint similar weaknesses in potential targets. More alarmingly, it is plausible that AI models, with their ability to analyze vast datasets of past exploits and known vulnerabilities through unsupervised learning methods, could accelerate the discovery of both known and novel weaknesses, including elusive zero-day threats.

Ironically, organizations can use the same AI-powered techniques to identify and address their own vulnerabilities. In other words, the means are the same. The tools are the same. The ends are different. The central issue then becomes which side will leverage AI models more effectively. Which side will be faster in detecting vulnerabilities? And when these vulnerabilities are found with the support of AI models, who will have first-mover advantage?

→ Recommendation G: Minimize external attack surface; for critical systems, strive for invisibility.

Given distributed assets, software-defined networks, and cloud computing, the idea of an "air-gapped" network is perhaps a thing of the past for most organizations and use cases. While the industry works to better enable adoption of zero trust architecture as a modernized approach to managing risk, ubiquitous AI-enabled network reconnaissance and probing described above invites a reprise of several old-fashioned concepts and controls.

Particularly when servicing essential critical infrastructure functions and handling particularly sensitive information, a defensible computer network must minimize, almost to the point of invisibility, public internet-exposed assets. It must also be effectively segmented and introduce strict controls at interfaces between IT and OT environments. Additional strategies for reducing an organization's discoverable attack surface are discussed in the Network Obfuscation section below.

Our Watch List

In the course of researching the topic of AI in cybersecurity and writing this report, the authors encountered several topics that are worthy of examination, but were either not ripe or somewhat adjacent to the central topic. This section memorializes these topics in short form, as a conversation starter and marker for potential future work.

Agentic AI Weaponization

As the capabilities of agentic AI continue to evolve, so too does the potential for misuse.[98] Some frontier labs are already employing AI agents that autonomously perform tasks, collaborate with other agents, and integrate seamlessly into workflows. This operational reality raises significant concerns about how these technologies could be weaponized.[99]

It is plausible that threat actors could employ AI agents to carry out an offensive cyber operation end-to-end, with little to no direct human supervision. Such agents could learn about target environments, develop attack strategies, select appropriate TTPs, and evaluate ways to increase plausible deniability. The AI agents could collaborate with human operators or even other AI agents throughout different stages of an operation. For instance, one agent might handle reconnaissance, while another focuses on resource collection. By delegating these tasks to AI agents, human operators could focus on other phases of the attack that require their attention and expertise.

One might also envision a future in which a multi-agent malicious environment flourishes. In this scenario, AI agents could train each other on malicious use cases, such as generating malware, breaching networks, or performing network or code obfuscation. This feedback loop would enable agents to continuously refine their skills, making them increasingly sophisticated. The potential for multi-agent collaborative learning could pose a formidable challenge for cybersecurity practitioners to identify, defend against, or attribute.

98 Tiernan Ray, "As AI agents spread, so do the risks, scholars say," *ZDNet*, March 4, 2024, https://www.zdnet.com/article/as-ai-agents-spread-so-do-the-risks-scholars-say/.

99 Jonathan Zittrain, "We Need to Control AI Agents," *The Atlantic*, July 2 2024, https://www.theatlantic.com/technology/archive/2024/07/ai-agents-safety-risks/678864/.

Code Obfuscation

Code obfuscation is a practice that long precedes the advent of AI tools, but AI models have certainly refined the practice, elevating code obfuscation to new levels of complexity and sophistication. A common and legitimate technique used to obscure code for intellectual property protection and to thwart reverse engineering, AI-enabled code obfuscation has now become an experimental tool of choice for adversaries aiming to evade Endpoint Detection and Response (EDR) systems and persist longer within compromised environments.[100,101] At the same time, AI provides powerful tools for decloaking these obfuscation techniques, offering new hope for enhancing threat detection and system security.

For malicious actors, code obfuscation adds layers of plausible deniability and serves as a tool for evasion and persistence. To complicate the defender's task of identifying malicious software, these actors can leverage code obfuscation techniques by renaming variables and functions, rearranging and breaking down executable code into confusing patterns, inserting redundant or misleading code, and using advanced algorithms to encrypt or morph code segments.

Cybercrime actors have long made use of third-party "crypting services to make their malware nearly invisible to antivirus software.[102] For instance, Cryptor[.]biz, a service allegedly active since 2016, has been known to provide crypting services to cybercriminals to obscure their malware. By the time Cryptor[.]biz was established, there was already a well-established demand and ecosystem for malware obfuscation services among bad actors.[103] In 2021, the Department of Justice found Koshkin, a Russian cyber criminal, providing, "malicious software fully undetectable by nearly every major provider of antivirus software."[104] This malicious service seemed to be effective across different types of malware, including "botnets, remote access trojans, credential stealers, cryptocurrency miners."[105]

More recently, researchers have found that AI models can automate the obfuscation process by editing or morphing malicious source code to conceal its origins and evade endpoint

100 Chris Brook, "What Is Code Obfuscation & How Does It Work?" Digital Guardian, May 2, 2024, https://www.digitalguardian.com/blog/what-code-obfuscation-how-does-it-work.

101 "Code Obfuscation: Protecting Your Software's Inner Workings," Verimatrix, August 28, 2023, https://www.verimatrix.com/cybersecurity/knowledge-base/the-power-of-code-obfuscation/.

102 "Operator of Counter Antivirus Service 'Scan4you' Sentenced to 14 Years in Prison," U.S. Department of Justice, September 21, 2018, https://www.justice.gov/opa/pr/operator-counter-antivirus-service-scan4you-sentenced-14-years-prison.

103 Brian Krebs, "Why Malware Crypting Services Deserve More Scrutiny," Krebs On Security (blog), June 21, 2023, https://krebsonsecurity.com/2023/06/why-malware-crypting-services-deserve-more-scrutiny/.

104 Krebs, "Why Malware Crypting Services Deserve More Scrutiny."

105 Krebs, "Why Malware Crypting Services Deserve More Scrutiny."

detection.[106] A recent report by OpenAI revealed that Crimson Sandstorm queried ChatGPT on different malware evasion techniques.[107] AI models also support various techniques like encryption, encoding, and polymorphic or metamorphic functions, which blend malicious code into benign patterns and evade detection systems.[108] These techniques complicate the identification of "known patterns of malicious activity [...] by blend[ing] irrelevant code into malware."[109] A 2023 research article also warned that AI-enabled code obfuscation for malware is increasingly concerning for current malware detection techniques and their ability to detect different strains of malware, since AI-enhanced "code obfuscation, code behavior adaptation, as well as learned communication detection evasion potentially bypass existing malware detection techniques."[110,111]

Code Deobfuscation

Organizations are experimenting with deobfuscation techniques to make the code easier to understand and analyze. Deobfuscation involves reversing obfuscation methods applied to software, restoring its structure, formatting, and naming conventions to a more understandable form. However, the effectiveness of current AI models in de-obfuscating malware code remains debated.

Though there are several challenges associated with using AI for code deobfuscation, we suspect that this frontier warrants further research and monitoring. Experts from Promon have identified several challenges associated with using AI for deobfuscation; the primary issues being model accuracy, adaptability, and data availability.[112] For ML models to effectively detect obfuscated code, they must first understand the context of the code being analyzed. While advancements in machine learning have improved models' general code parsing and interpretation abilities, recognizing obfuscated code remains a significant challenge. Techniques such as random variable renaming, insertion of irrelevant code or data, and encryption or compression make it challenging for a model to pinpoint and learn consistent patterns.

106 Christine Barry, "5 Ways Cybercriminals Are Using AI: Malware Generation," Barracuda, April 16, 2024, https://blog.barracuda.com/2024/04/16/5-ways-cybercriminals-are-using-ai--malware-generation.

107 OpenAI, "Disrupting Malicious Uses of AI."

108 Rahul Awati, "What Are Metamorphic and Polymorphic Malware?" *TechTarget*, March 2022, https://www.techtarget.com/searchsecurity/definition/metamorphic-and-polymorphic-malware.

109 Awati, "What Are Metamorphic and Polymorphic Malware?"

110 Lothar Fritsch, Aws Jaber, and Anis Yazidi, "An Overview of Artificial Intelligence Used in Malware," *Communications in Computer and Information Science* 1650 (2022): 41–51, https://doi.org/10.1007/978-3-031-17030-0_4.

111 Fritsch, Jaber, and Yazidi, "An Overview of Artificial Intelligence Used in Malware."

112 Dr. Anton Tkachenko, "AI Deobfuscators: Why AI Won't Help Hackers Deobfuscate Code (Yet)," Promon, May 8, 2024, https://promon.co/security-news/ai-deobfuscator-hackers-deobfuscate-code.

The quality and availability of data repositories to train these models to deobfuscate malicious code is also crucial. Effective model training requires datasets that include both obfuscated and unobfuscated software examples to expose the model to a wide variety of obfuscation techniques. Furthermore, the data must be accurately labeled to allow the model to differentiate between altered and original code states, a task that often demands substantial manual effort. AI systems currently lack the nuanced understanding that human experts possess, making this labeling process critical.[113] Thus, model capabilities and the availability of high-quality training data are two significant hurdles in applying AI for code deobfuscation.

However, other researchers argue that current models, such as OpenAI's ChatGPT, are making significant strides in code deobfuscation, especially in supporting organizations with a strong understanding in analyzing obfuscated code. A recent article posits that, "OpenAI's language model excels (sometimes more, sometimes less) in de-obfuscating script-based malware."[114]

Since malware analysis often demands a specific set of tools and know-how, organizations are allegedly exploring the integration of OpenAI's plugins into malware analysis tools like Ghira, which would then generate context-specific outputs about the decompiled code. These plugins, such as GPTHidra and G-3PO, are accelerating the analysis of obfuscated code.[115,116] Researchers are optimistic that, as AI model performance improves, the automation of code deobfuscation may become more effective and accurate, enhancing organizations' ability to quickly "understand more complex code and identify vulnerabilities."[117]

Polymorphic Malware and Evasion

More recently, both industry and academic circles have flagged the rise of AI-enabled polymorphic malware.[118] Polymorphic malware continuously morphs its code with each execution, rendering traditional detection methods, such as signature-based methods (which flags a threat based on known patterns or signatures in the malware's code), ineffective.[119] This malware can change file names, types, and internal code structures, even encrypting parts of itself with different keys in each iteration. The constant evolution and variations of its digital

113 Andreas Holzinger et al., "Toward Human-Level Concept Learning: Pattern Benchmarking for AI Algorithms," *CellPress* 4, no. 8 (July 1, 2023): 100788–88, https://doi.org/10.1016/j.patter.2023.100788.

114 Alessio Trivisonno, "Reverse Engineering : Code Deobfuscation in the Age of AI," Medium, June 27, 2023, https://infosecwriteups.com/the-cybersecurity-revolution-at-the-age-of-ai-openai-and-code-deobfuscation-3f9dd278b900.

115 Evyatar9, "GptHidra," GitHub repository, accessed August 3, 2024, https://github.com/evyatar9/GptHidra.

116 Tenable, "Ghidra Tools, G-3PO: A Protocol Droid for Ghidra," GitHub repository, accessed August 3, 2024, https://github.com/tenable/ghidra_tools/tree/main/g3po.

117 Trivisonno, "Reverse Engineering."

118 Dena De Angelo, "The Dark Side of AI in Cybersecurity — AI-Generated Malware," Palo Alto Networks, May 15, 2024, https://www.paloaltonetworks.com/blog/2024/05/ai-generated-malware/.

119 De Angelo, "The Dark Side of AI in Cybersecurity."

footprint make tracking and containment exceedingly difficult. This trend, as highlighted by Palo Alto Networks researcher Rem Dudas, involves inputting snippets of malware source code into an LLM. The goal is to generate a range of malware samples that retain the same core functionality, but vary in their source code. Although these samples perform similar malicious functions, the diversity in their code and the sheer volume of unique variants can overwhelm the capacity of detection and analysis teams, making it significantly more challenging to identify and counteract.

The characteristics of polymorphic malware—its ability to mutate, encrypt, and obfuscate—means that each instance can alter its code during the replication or infection processes. This not only hides the payload through encryption but also disguises its true functionality using techniques like dead code insertion, register renaming, and instruction substitution.

Two notable proofs of concept demonstrate the utility and sophistication of AI-enabled malware: BlackMamba and Deeplocker.

Dubbed BlackMamba, AI-generated polymorphic malware has already been proven to bypass industry leading EDR, a proof of concept project led by HYAS Labs researchers.[120,121] The attack "exploits a large language model to synthesize a polymorphic keylogger functionality on the fly." The malware is described as "'truly polymorphic' in that every time BlackMamba executes, it resynthesizes its keylogging capability. [It] demonstrates how AI can allow the malware to dynamically modify benign code at runtime without any command-and-control (C2) infrastructure, allowing it to slip past current automated security systems that are attuned to look out for this type of behavior to detect attacks."[122]

DeepLocker, another prominent example of AI-enabled malware evasion, conceals its malicious functions within seemingly innocuous applications, such as video conferencing software.[123] The core innovation of DeepLocker lies in its use of AI to establish highly specific "trigger conditions" for activating its payload, given that the malware appears harmless until triggered; conditions designed to be extremely difficult to reverse engineer. The malware remains dormant until it identifies its target using a DNN AI model trained to activate only under certain criteria, such as specific visual, audio, geolocation, or system-level characteristics. DeepLocker's AI model functions as a sophisticated lock, with the trigger

120 Jeff Sims, "BlackMamba: Using AI to Generate Polymorphic Malware," Hyas, July 31, 2023, https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware.

121 Migo Kedem, "BlackMamba ChatGPT Polymorphic Malware: A Case of Scareware or a Wake-up Call for Cyber Security?" SentinelOne, March 16, 2023, https://www.sentinelone.com/blog/blackmamba-chatgpt-polymorphic-malware-a-case-of-scareware-or-a-wake-up-call-for-cyber-security/.

122 Elizabeth Montalbano, "AI-Powered 'BlackMamba' Keylogging Attack Evades Modern EDR Security," *Dark Reading*, March 8, 2023, https://www.darkreading.com/endpoint-security/ai-blackmamba-keylogging-edr-security.

123 Marc Stoecklin, Jiyong Jang, and Dhilung Kirat, "DeepLocker: How AI Can Power a Stealthy New Breed of Malware," Security Intelligence, August 8, 2018, https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/.

conditions serving as the key. This design makes DeepLocker exceptionally stealthy, with its complex DNN making it difficult for analysts to dissect or predict its behavior.

BlackMamba and Deeplocker illustrate that developing AI-enabled malware proofs of concept is not out of reach. While current open source evidence does not indicate widespread use of AI for malware generation by adversaries, the potential for such threats looms large. As one respondent warned, "[Although novel capabilities only exist within the] proof of concept stages rather than the actual deployment stages, [I] would be surprised they [don't exist] and shocked if that kind of thing isn't happening. Someone has proven the concept somewhere whether they're using it or not."

Network Obfuscation

Similar in purpose to code obfuscation and related to the AI-enabled botnet example above, *network* obfuscation is a cybersecurity strategy that may be employed by enterprise owners to hide network assets and data-in-transit from bad actors. Retired CIA technical expert Barbara Hunt explained, "Many enterprises mistakenly assume that their wide-area network (WAN) is inherently secure when, in fact, any network segment or connection is a potential source of attack....By disguising and controlling change across multiple system dimensions, network obfuscation increases uncertainty and complexity for attackers, reducing their window of opportunity and increasing the costs of their probing and attack efforts."[124] Cybersecurity vendor SecureCo further explained that "network obfuscation can take many forms, but is generally accomplished through technology strategies that employ stealth, evasion and anonymization" through, for example, dynamic routing and endpoint concealment, thereby reducing the extent of an organization's discoverable attack surface.[125] Ms. Hunt furthermore described dynamic routing as a "moving-target" defense technique.

Malicious cyber actors have also long used network obfuscation techniques to route and launder their traffic so as to conceal its true source and make it harder to detect and defend against. In the early days of so-called Advanced Persistent Threat (APT) activity, such networks were merely compromised small business computer systems, commonly referred to as "hop points" or "operational relay boxes." This tradecraft has since evolved to include the use of Infrastructure as a Service (IaaS) products to obfuscate foreign-based malicious traffic by appearing as domestic in origin and evade government surveillance by rapidly provisioning,

124 Barbara Hunt, "Understanding The Power Of Network Obfuscation," *Forbes*, September 29, 2021, https://www.forbes.com/councils/forbestechcouncil/2021/09/29/understanding-the-power-of-network-obfuscation/.

125 "Network Obfuscation: The Key to Protecting Your Company's Network and Data," SecureCo, June 30, 2023, https://www.secureco.com/posts/network-obfuscation-key-to-protecting-your-network/.

using, and abandoning accounts before they can be investigated.[126] Additionally, the U.S. Department of Justice and Federal Bureau of Investigation this year took action to disable a botnet consisting of hundreds of small office/home office routers used by the Russian military to route and conceal their cyber operations.[127] Increasingly, this operational infrastructure—whether it involve virtual private servers (VPS), compromised routers, or IoT devices—is being referred to as "obfuscation networks." Identifying, observing, and disrupting this infrastructure and the operations it facilitates should be a key goal of responsible states' law enforcement and intelligence services.

WHY IS THIS TOPIC RELEVANT TO AI?

While somewhat speculative, this report posits that AI will soon play a role in bad actors' means of operating and maintaining their obfuscation infrastructure, but also assist defenders in uncloaking it. An industry representative interviewed for this report explained his company's experience with LLMs being immediately useful in performing defined tasks, such as producing boilerplate code for defining infrastructure in a cloud environment. Consider a future in which malicious actors employ AI to achieve stealth by spinning up new VPSs, introducing uniqueness to each configuration; and dynamically routing attack traffic.

Conversely, in which key ecosystem enablers like cloud providers, telecommunication companies, and cybersecurity vendors could join forces—potentially making use of a federated learning platform—to dynamically teach the ecosystem to identify bad actors' obfuscation nodes and networks. This could enable coordinated efforts to deny their use of the infrastructure, pushing them to less desirable platforms and hampering their malicious activities.

126 White House, Executive Order 13984, 86 FR 6837 (January 19, 2021), https://www.federalregister.gov/documents/2021/01/25/2021-01714/taking-additional-steps-to-address-the-national-emergency-with-respect-to-significant-malicious.

127 "Justice Department Conducts Court-Authorized Disruption of Botnet Controlled by the Russian Federation's Main Intelligence Directorate of the General Staff (GRU)," press release, Office of Public Affairs, U.S. Department of Justice, February 15, 2024, https://www.justice.gov/opa/pr/justice-department-conducts-court-authorized-disruption-botnet-controlled-russian.

Conclusion

Near-term, the AI in cybersecurity advantage goes to the defender. The "home field" advantage—which includes access to proprietary software source code (for the software maker), a full understanding of network architecture and typical user patterns (for the enterprise network owner), and an ecosystem of service providers who are making rapid strides to capitalize on the potential of AI—will be difficult for an adversary to overcome. Furthermore, first-mover advantage seems to be squarely with western and likeminded governments and technology firms.

Additionally, the cyber workforce is already realizing the benefits of AI across numerous arduous tasks that have not traditionally scaled well. This includes writing more secure code, finding and fixing bugs and flaws in large codebases and long-neglected open source software libraries, and re-writing software in memory safe languages. (This may be the surest path to achieving "secure by design" goals.) Immediate benefits are also being realized in security operations, enabling SOC analysts to be more efficient, freeing them up to perform higher-level tasks, and potentially improving their job satisfaction.

However, opportunities for bad actors are also quite significant, but as of the date of this writing, only the most sophisticated state actors are likely keeping pace. Causes for concern include bad actors' use of LLMs in analyzing, summarizing, and generating content for use in their attack cycle. Generative AI's application in personalized, context-rich phishing and impersonation is quite compelling. This proven superpower is available to actors of all stripes, from the lowliest ransomware gang to the "pacing threat" state actor.

While the longer-term outlook is uncertain and might be dismal for poorly defended enterprises, by fully capitalizing on first-mover advantage, key ecosystem enablers and organizations they serve might establish a defensible posture that is prepared for whatever the future might bring. Staying ahead will require continued investment, innovation, and integration, as this is an arms race that is just getting started.

INSTITUTE FOR SECURITY AND TECHNOLOGY
www.securityandtechnology.org

info@securityandtechnology.org