CYBER THREAT ALLIANCE

2025

CYBERSECURITY IN THE AGE OF GENERATIVE AI:

PART I

Combating GenAI Assisted Cyber Threats

CYBER THREAT ALLIANCE

POWERED BY THE CTA

The Cyber Threat Alliance (CTA) is the industry's first formally organized group of cybersecurity practitioners who work together in good faith to share threat information and improve global defenses against cyber adversaries. CTA facilitates the sharing of cyber threat intelligence to improve defenses, advance the security of critical infrastructure, and increase the security, integrity, and availability of IT systems.

We take a three-pronged approach to this mission:

1. Protect End-Users: Our automated platform empowers members to share, validate, and deploy actionable threat intelligence to their customers in near-real-time.

2. Disrupt Malicious Actors: We share threat intelligence to reduce the effectiveness of malicious actors' tools and infrastructure.

3. Elevate Overall Security: We share intelligence to improve our members' abilities to respond to cyber incidents and increase end-user's resilience.

CTA is continuing to grow globally, enriching both the quantity and quality of the information shared among its membership. CTA is actively recruiting additional cybersecurity providers to enhance our information sharing and operation collaboration to enable a more secure future for all.

For more information about the Cyber Threat Alliance, please visit:
https://cyberthreatalliance.org.

CYBERSECURITY IN THE AGE OF GENERATIVE AI WORKING COMMITTEE MEMBERS

Check Point Software Technologies Ltd.
Amit Sharon
Sergey Shykevich

CyberCX
Mark Hofman

Defending Digital Campaigns
Tiffany Schoenike

Fortinet
Val Saengphaibul
James Slaughter

FS-ISAC
Michael Silverman
Dr. Carrie E. Gates

H-ISAC
Ethan Muntz

IT-ISAC
Ian Andriechack

Juniper Networks
Asher Langton

McAfee
Christy Crimmins
Abhishek Karnik
German Lancioni

NGO-ISAC
Ian Gottesman

NTT
David Beabout

Rapid7
Laura Ellis

Scitum
Imelda Flores

Symantec by Broadcom
Scott Swett
Brian Ewell

Cyber Threat Alliance
Chelsea Conard
Michael Daniel
Jeannette Jarvis
Linda Beverly
Kate Holseberg

This report also leverages shared data and published analysis from Palo Alto Networks and Gen Digital. CTA members reviewed the document and the report reflects our shared consensus.

This report would not be possible without the generous support from Craig Newmark Philanthropies.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This Joint Analytic Report (JAR), *Cybersecurity in the Age of Generative AI: Part I — Combating GenAI Assisted Cyber Threats*, investigates the ways in which adversaries leverage Generative AI (GenAI), such as Large Language Models (LLMs) and related technologies like those enabling deepfakes, to create AI-assisted threats. This report is accompanied by a Part II that extends the analysis to encompass threats targeting the broader AI Ecosystem.

This JAR provides an overview of adversaries' current use cases for AI-assisted threats and debunks common misconceptions about GenAI's role in cybersecurity. The findings highlight that GenAI enhances malicious activity across mediums including text, audio, images, and video. Adversaries are employing techniques such as phishing, malware development, deepfakes, and adaptive threats to enable more efficient and scalable campaigns. However, despite the media hype, this JAR emphasizes that GenAI is not making the adversary smarter; it is making the adversary more efficient.

This report underscores the need for organizations to integrate traditional cybersecurity practices with GenAI-specific policies to strengthen their defenses. By adopting a proactive approach, organizations can mitigate the risks posed by AI-assisted threats during the early stages of malicious actors leveraging AI tools. However, organizations should act now because this window may not stay open for long.

INTRODUCTION

In November 2022, ChatGPT seemingly exploded into public awareness and brought the term *Generative Artificial Intelligence* (GenAI) into the mainstream. Although various forms of *Artificial Intelligence* (AI) had been in use for years, ChatGPT made AI accessible to the general public in a different way and it sparked a revolution in the use of that technology. In the two years since ChatGPT's debut, many claims have been made about the technology's capacity to create benefits and harms, including in cybersecurity. Much of this language was couched in apocalyptic terms or as the savior of all, with extreme claims about AI's capabilities.

However, in many cases the hype has far outstripped reality. While AI in general and GenAI in particular will change many aspects of cybersecurity, the effects so far have been more limited than anticipated and the projected impacts will likely take longer to realize than many analysts originally projected. At the same time, adversaries are using GenAI to augment their capabilities and we are already seeing some of the effects of that adoption. The real question is how are adversaries actually using GenAI tools and what can we reasonably expect over the next six months to a year?

This JAR tries to answer those questions based on the Cyber Threat Alliance's (CTA) collective expertise. It explores how adversaries leverage tools like GPT to manipulate GenAI and LLMs to craft AI-assisted threats for malicious purposes, including employing other AI technologies, such as those behind deepfakes, for additional manipulation. This report bases its analysis on data and evidence-based case studies available to CTA members, and it dispels common myths about GenAI as an entirely new threat vector. The result is sobering rather than sensational. We hope that readers will incorporate insights from this work into their cybersecurity planning, taking a measured approach that adapts to the changes GenAI will bring to the landscape without panicking or overreacting. This JAR also offers guidance on how to respond, providing actionable steps that creators and users of the technology can apply to protect and defend themselves. GenAI will indeed bring about many changes, but it may be different than what we think or fear.

What is Artificial Intelligence?

The term *Artificial Intelligence* (AI) encompasses a broad array of technologies. The AI Ecosystem is composed of different types of AI. At its foundation lies the *AI Infrastructure*, which includes the computer hardware and data pipelines necessary to process large amounts of data. Built upon this infrastructure are different types of AI models, each with varying capabilities: For example, *predictive AI* uses historical and real-time data to forecast future events or behaviors, and *prescriptive AI* recommends specific actions.[1] These types of AI are often associated with traditional approaches like machine learning and neural networks.

Generative AI (GenAI) introduces an additional dimension by enabling the creation of new data. Thus, GenAI is a class of AI that not only interprets, but also generates original content, advancing the potential of data analysis and innovation. *Agentic AI* is connected to systems and *Application Programming Interfaces* (APIs) can perform actions without human intervention.[2]

GenAI relies on mathematical algorithms capable of generating entirely new content, such as faces or text. For example, *Large Language Models* (LLMs) leverage specific algorithms to generate new text based on the

[1] https://one.oecd.org/document/GOV/SBO(2024)14/en/pdf

[2] Unless explicitly cited, the definitions are derived from the Working Committee discussions held between August 22 and November 7, 2024. These definitions reflect the consensus of the Working Group and are available in the Definitions section at the end of this report.

content that has been provided as input. Other GenAI approaches use different algorithms to generate new images and faces. To make these models accessible and user-friendly, companies rely on technologies like *Generative Pre-trained Transformers* (GPTs) to power interfaces and applications. We include *deepfakes*[3] in this generic category, while recognizing that not all deepfake generation processes necessarily use GenAI algorithms (e.g., voice cloning for audio deepfakes may leverage vocoders and speech synthesizers rather than GenAI algorithms).

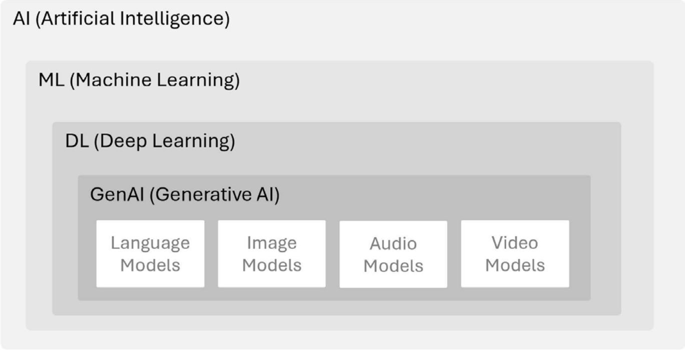

Image provided by German Lancioni

The definitions section at the end of this report includes some examples of algorithms and techniques used with GenAI, such as machine learning, GPT models, deep learning methods, and Generative Adversarial Networks (GANs).

AI-ASSISTED THREATS

Unquestionably, malicious cyber actors are using GenAI as part of their activities. The issue is how they are incorporating GenAI into their operations. Speculation about what adversaries *might* do has run rampant, and researchers have developed a large number of "proofs of concept." However, a proof of concept demonstrates that a technique or approach *could* work; the existence of a proof of concept does not automatically mean that a malicious actor is using the technique.

This combination of speculation and research has created a lot of confusion about what malicious actors are actually doing with GenAI, because many articles and reports treat speculation, proofs of concept, and data-driven analysis in the same manner.

In order to reduce the confusion, this paper focuses on data-driven analysis: where do we have data that shows adversaries are using AI? This data can be direct (for example, an AI-generated audio used in a scam) or indirect (an increase in the volume of phishing emails, which could result from using GenAI to create text more efficiently). To help structure our analysis, we first discuss the mediums in which malicious actors can use AI, then we turn to the techniques for GenAI use, followed by the purpose for using GenAI. In this work, we focus on *AI-Assisted Threats*, which are cases where AI models are used as tools to assist malicious actors to generate content for malicious purposes. The accompanying Part II to this JAR examines *Adversarial Threats*, where malicious actors target AI models themselves, addressing broader risks to the AI ecosystem.

MEDIUMS

Broadly speaking, malicious actors can use GenAI in four mediums: text, audio, video, and images. None of these mediums represent new channels for malicious cyber activity, but GenAI makes working in these mediums easier and faster.

Text

GenAI can be used to produce text. Just as everyday users use AI to draft text in emails and social media, malicious actors are also leveraging AI tools to craft

[3] A deepfake is synthetic media that has been digitally manipulated, including audio, images, and video. (Extended definition provided in the definition section of this report.)

highly convincing emails, messages, and website content. These text-based threats span various tactics, including phishing,[4] smishing (SMS phishing),[5] and other forms of digital deception.

Audio

Audio clones, or synthetically generated voice recordings, can replicate natural voices with remarkable accuracy. While they have legitimate applications, such as recreating an actor's voice for media or narrating e-books with consent, malicious actors are now using GenAI to produce audio clones for deceptive purposes. These clones make social engineering more convincing and enable the creation of realistic fake profiles on social media. The sophistication of these audio clones complicates users' ability to distinguish genuine and fraudulent communication.

Images

Malicious actors are also using GenAI to create two types of deceptive images: fully generated images and inpainted images. *Fully generated images* are created entirely from scratch by a GenAI tool, and *inpainted images* begin with a real photo and are subtly altered by AI. Detecting inpainted images can be particularly difficult because the modifications are often inconspicuous. This challenge grows as image manipulation becomes more prevalent in society; for instance, modern smartphones offer AI-driven adjustments such as skin smoothing or enhanced backgrounds that often go unnoticed. Even a seemingly simple action like removing an individual from a background can technically qualify as a deepfake image creation. The widespread availability of these tools raises important questions about what constitutes acceptable image manipulation versus alterations that should be flagged for malicious use.

Video

Similar to images, GenAI can be used to either modify existing videos or create entirely new ones. The quality and realism of these videos enable malicious actors to deceive or mislead users by impersonating individuals or fabricating events.

TECHNIQUES

To exploit GenAI across these various mediums, malicious actors are leveraging the technology in both isolated cases and broader, recurring campaigns. Less frequent, isolated instances demonstrate the versatility of AI in cybersecurity threats. For instance, researchers have been able to use GenAI to obfuscate malware signatures, which would complicate detection by traditional security systems; however, no CTA members have identified malware "in the wild" as being AI-generated.[6] More commonly, malicious actors use GenAI to enhance social engineering, assist in malware authoring, optimize command and control operations, or produce deepfake media.

Social Engineering

Cybercriminals can now use both publicly available and illicitly managed, specialized AI tools to craft highly convincing phishing emails and SMS messages that match the style and tone of the impersonated entity. These tools generate error-free, professional text that is difficult for recipients to distinguish from human-written content, often bypassing both human

[4] Phishing involves sending emails with malicious links or attachments under the guise of legitimate sources, targeting a broad audience without much personalization (https://www.cyberthreatalliance.org/resources/assets/cyber-threats-to-ngos/)

[5] Smishing utilizes SMS text messages to trick recipients into revealing personal information by posing as reputable entities, exploiting the immediacy and personal nature of text messages (https://www.cyberthreatalliance.org/resources/assets/cyber-threats-to-ngos/)

[6] https://www.paloaltonetworks.com/blog/2024/05/ai-generated-malware/

scrutiny and automated email filters.[7] Additionally, Gen-AI powered phishing messages can be written more easily in multiple languages, expanding the number of potential targets for malicious actors. Thus, GenAI tools make creating phishing emails much easier and faster, as well as improving their quality; according to New Scientist, GenAI-generated phishing emails can be up to 96% more cost efficient than those written by humans.[8] Supporting this efficiency assertion, some evidence indicates a 1,265% increase in the volume of phishing emails since ChatGPT's launch in late 2022.[9] McAfee Labs has run laboratory experiments indicating that current generation LLMs could increase the success or "click" rate from about 20% to 30%.[10] However, CTA members do not currently have data from clients and customers demonstrating that AI-generated phishing emails have a higher click rate compared to non-AI generated phishing emails.

Malware

Malicious actors can use GenAI in the same way legitimate software engineers use AI tools to enhance their productivity and efficiency.[11] By using publicly available tools, threat actors can streamline the creation of malware, customizing their code at a faster rate than previously possible. Unlike traditional malware, AI-generated malware combines machine-driven code with human oversight, allowing developers to tailor the final product for specific malicious purposes. The combination of AI and human expertise makes AI-generated malware hard to detect. For example, in December 2022, a hacker used ChatGPT to replicate malware strains, including a Python-based infostealer that targeted common file types, copied them to a random folder within the Temp directory, compressed them into a ZIP file, and uploaded them to an FTP server.[12] This malware, however, lacked encryption or secure transfer, risking exposure of the stolen files to unintended third parties. In another case from 2024, the phishing campaign "CopyRh(ight)adamantys" used AI-supported tools to craft multilingual phishing emails that impersonated media and tech companies in order to deploy malware that targeted victims globally.[13] OpenAI has confirmed that nation-state actors have exploited ChatGPT to accelerate malware development.[14] However, while these examples show what is possible, CTA members and partners do not have evidence of widespread malware generation by AI tools.

Command and Control

GenAI tools are also enhancing the ability of malicious actors to manage large networks of compromised devices or fake social media accounts; this management process is often referred to as "command and control." For example, botnets are networks of infected devices controlled remotely by a malicious actor without the consent and knowledge of the devices' owners,[15] and they can be used to support Distributed Denial of Service (DDoS) attacks. A DDoS attack overwhelms a targeted online service with a large volume of requests, exhausting system resources and rendering the services unavailable.[16] Malicious actors can use GenAI tools to manage these

[7] https://arstechnica.com/information-technology/2023/07/why-ai-detectors-think-the-us-constitution-was-written-by-ai/

[8] https://www.newscientist.com/article/2361490-chatgpt-can-be-made-to-write-scam-emails-and-it-slashes-their-cost/

[9] https://slashnext.com/press-release/slashnexts-2023-state-of-phishing-report-reveals-a-1265-increase-in-phishing-emails-since-the-launch-of-chatgpt-in-november-2022-signaling-a-new-era-of-cybercrime-fueled-by-generative-ai/

[10] https://www.mcafee.com/blogs/other-blogs/mcafee-labs/the-dark-side-of-gen-ai/

[11] https://www.expresscomputer.in/amp/artificial-intelligence-ai/hp-threat-researchers-uncover-evidence-of-attackers-using-ai-to-generate-malware/116517/

[12] https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/

[13] https://research.checkpoint.com/2024/massive-phishing-campaign-deploys-latest-rhadamanthys-version/

[14] https://www.bleepingcomputer.com/news/security/openai-confirms-threat-actors-use-chatgpt-to-write-malware

[15] https://www.cyberthreatalliance.org/resources/assets/cyber-threats-to-ngos/

[16] https://www.cyberthreatalliance.org/resources/assets/cyber-threats-to-ngos/

botnets more efficiently.

Deepfake

A deepfake is synthetic media that has been digitally manipulated to create false content. Deepfakes leverage techniques from machine learning and AI to manipulate or generate visual and audio content to imitate a real-world equivalent, often with the purpose of deceiving the audience into believing its authenticity. Although this kind of manipulation existed prior to the availability of GenAI tools, these tools make generating such content easier and improve the overall quality of the output. This technique may be one of the most prevalent uses of GenAI tools.

HOW ARE MALICIOUS ACTORS USING GENAI TOOLS?

Rather than using AI to conduct novel activities, malicious actors have primarily used GenAI tools to make the kinds of activities they were already doing more efficient or effective. Thus, malicious actors are using GenAI for cybercrime activities like scams or fraud, mis- and disinformation campaigns, and personal threats, such as reputational attacks or bullying. In this context, GenAI becomes either a “force multiplier” (making the activity more effective) or a “participation enabler” (making it easier for more people to engage in the activity).

Leveraging GenAI for Scams

Although malicious actors are using GenAI to make text-based phishing more effective, the most innovative way malicious actors leverage GenAI for scams is through deepfakes. While the term “deepfake” typically brings to mind video manipulation, audio deepfakes may pose an even greater threat, particularly in financial fraud. Audio clones, or synthetically generated voice recordings, are becoming increasingly sophisticated. In the financial services sector, for example, malicious actors will impersonate a bank customer using an audio clone, bypassing traditional security checks and convince staff to approve fraudulent transactions.[17]

GenAI has also transformed traditional scams into more convincing schemes by leveraging deepfake technology in various environments. Malicious actors will replicate a victim’s loved one’s voice in an audio scam to manipulate him/her into transferring money urgently. In these cases, a victim will receive a distressing call from someone who sounds exactly like his/her child or grandchild pleading for help.[18] This emotional manipulation exploits the inability to reliably detect AI-generated audio. While regulation and consent requirements may force the implementation of controls that may hinder the practice, it is unlikely that threat actors will take heed unless it starts becoming ineffective or expensive. Similarly, investment scams use AI-generated voices or videos to impersonate financial advisors and celebrities, encouraging victims to invest in fraudulent opportunities that span retirement funds, ponzi schemes, and cryptocurrency scams.[19]

In a related tactic, a deepfake video was used in at least one case to impersonate a CEO, fraudulently authorizing a transaction by convincing employees they received legitimate instructions.[20] Scammers are also using deepfake technology to bolster online dating scams, using realistic images and videos to build trust with their targets.[21] In a recent case, a Hong Kong crime ring swindled victims out of $46 million through elaborate fake identities on

[17] https://www.fsisac.com/newsroom/deepfake-technology-poses-new-threats-to-financial-institutions-fsisac-provides-guidance

[18] https://www.crn.com/news/ai/2024/audio-deepfake-attacks-widespread-and-only-going-to-get-worse

[19] Sensity AI. The State of Deepfakes 2024. Available for download at https://sensity.ai/reports/

[20] https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

[21] https://media.mcafeeassets.com/content/dam/npcld/ecommerce/en-us/docs/reports/rp-mcafee-modernlove-report.pdf

dating platforms.[22] Meanwhile, video scams are also featuring celebrities to endorse products and mislead viewers into making purchases under false pretenses. [23]

Although GenAI tools are probably not responsible for the total increase in fraud over the past two years, these tools have likely contributed to a significant rise in investment fraud losses. According to the FBI, investment fraud losses increased 38% from $3.31 billion in 2022 to $4.57 billion in 2023.[24] According to Deloitte, their value for AI fraud risk is higher, sitting at $12.3 billion in 2023 in the United States.[25] This increase underscores the growing need for regulatory oversight and public awareness of AI-driven scams.

Leveraging GenAI for Mis- and Disinformation

In addition to fraud and other cybercrime, malicious actors can use GenAI tools to create, distribute, and amplify mis- and disinformation. This activity can serve political purposes, economic motives, or even personal vendettas. The deepfake technique is the most prevalent in this type of malicious activity.

The democratization of AI-driven editing tools has also lowered the barriers to create falsified content, enabling not only state actors and organized criminals, but also ordinary users to manipulate images and videos. Unlike political disinformation campaigns, which are often driven by nation-state actors, this type of activity focuses on creating or amplifying economic misinformation, public health conspiracies (such as vaccine-related falsehoods),[26] or other socially related narratives. Unfortunately, the speed of such content on social platforms makes it nearly impossible to pull down or counter the false narratives adequately.

According to Check Point Research, deepfakes are becoming part of election campaigns worldwide, employed by political candidates for self-promotion, defamation of opponents, or, more concerningly, as instruments for surreptitious foreign interference to destabilize democratic institutions.[27] This technique has been observed in multiple regions, including Argentina, Slovakia, and India, where fabricated content has shown up in political materials[28] and in Latin America to create false narratives by impersonating leaders.[29] Further, nation-state actors are leveraging GenAI tools, like ChatGPT, to amplify disinformation campaigns, such as in the recent U.S. presidential race.[30] In Pakistan, AI tools were used to create a victory speech attributed to the former Prime Minister while he was in jail for the purpose of amplifying political messaging and misleading the public.[31] These disinformation campaigns are often timed strategically, launching just before elections to exploit legal gaps, like media blackouts, which limit the time for fact-checking and public clarification.[32]

Deepfakes can also alter historical records by creating altered images or videos designed to skew public perception and bolster propaganda. This technique is not new, historical examples include Soviet

[22] https://www.darkreading.com/cyberattacks-data-breaches/hong-kong-crime-ring-swindles-victims-out-of-46m

[23] https://www.bbb.org/article/scams/18549-scam-alert-con-artists-impersonate-your-favorite-celebrity

[24] https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf

[25] https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2024/deepfake-banking-fraud-risk-on-the-rise.html

[26] https://www.bu.edu/ceid/2024/04/25/how-can-we-tackle-ai-fueled-misinformation-and-disinformation-in-public-health/

[27] https://research.checkpoint.com/2024/beyond-imagining-how-ai-is-actively-used-in-election-campaigns-around-the-world/

[28] https://research.checkpoint.com/2024/beyond-imagining-how-ai-is-actively-used-in-election-campaigns-around-the-world/

[29] https://www.axios.com/2024/02/22/misinformation-generative-ai-deepfakes-spanish-language

[30] https://www.npr.org/2024/08/17/nx-s1-5079397/openai-chatgpt-iranian-group-us-election

[31] https://www.politico.eu/article/pakistans-imran-khan-use-ai-artificial-intelligence-make-victory-speech-from-jail/

[32] https://research.checkpoint.com/2024/beyond-imagining-how-ai-is-actively-used-in-election-campaigns-around-the-world/

alterations of photographs as early as the 1930s[33]; however, GenAI tools have made creating such media much easier. For instance, in November 2024, a deepfake video falsely depicted civil rights leader Martin Luther King Jr. posthumously endorsing a political candidate.[34] The growing availability of deepfake tools, often accessible at minimal cost through open-source platforms or dark web markets, has made this technology an increasingly potent weapon in disinformation campaigns.

Additionally, the recently exposed "Green Cicada Network" by CyberCX demonstrates the potency of GenAI tooling. An AI-controlled network of inauthentic social media accounts was created to converse with other users. This network, comprising over 5,000 inauthentic accounts, was being tested to engage users on a number of topics and actively spreading divisive political messages on platforms like X (formerly Twitter), in democracies like the U.S. and Australia.[35] While this network was dismantled before it was fully operational, it demonstrated that a small number of individuals can potentially influence a large population using GenAI.

Leveraging GenAI for Direct Personal and Reputational Harms

The accessibility of GenAI has made it possible for individuals not traditionally considered “malicious actors” to use this technology to harm others. Non-technical users are now creating deepfakes to damage reputations, both personal and professional. In one case, a U.S. school employee fabricated racist audio and falsely attributed it to a colleague, causing significant personal and professional harm.[36] Deepfakes have also been used to insert celebrities and private individuals into pornographic content, such as a recent case involving Taylor Swift, where AI-generated explicit images of her circulated online.[37]

As deepfakes expand beyond celebrity targets to personal threats on everyday individuals, the risks of harm extend to both children and adults, amplifying concerns over privacy and online safety. Therefore, while political entities may leverage deepfakes for broader manipulation, individuals can also use GenAI for personal vendettas, amplifying the impact on private lives and households in society. As a result, deepfake technology is also being used for bullying. Some applications allow users to digitally remove clothes from an individual’s image, a disturbing trend that often targets students and has raised alarms in schools.[38]

DEBUNKING MYTHS

Although the previous section provides multiple examples of the way malicious actors are using GenAI tools, they are more anecdotal than systemic and more evolutionary than revolutionary. Thus, some frequently repeated assertions about the impact of GenAI on cybersecurity are not supported by the evidence. This section will address some of the most common myths surrounding GenAI and clarify its actual role and risks in the cybersecurity landscape.

“AI HAS TRANSFORMED CYBER THREATS”

A common misconception is that recent advancements in AI represent a revolution in cyber threats. While AI innovations are undeniably powerful, they have so far resulted in incremental

[33] https://www.history.com/news/josef-stalin-great-purge-photo-retouching

[34] https://www.independent.co.uk/news/world/americas/us-politics/martin-luther-king-trump-deepfake-b2641309.html

[35] https://cybercx.com.au/blog/cybercx-unmasks-china-linked-ai-disinformation-capability/

[36] https://forward.com/fast-forward/606376/former-employee-used-ai-to-frame-baltimore-school-principal-as-antisemitic-and-racist-police-say/

[37] https://www.axios.com/2024/02/03/taylor-swift-deepfake-ai-image-protection

[38] https://www.pcmag.com/news/apple-pulls-3-generative-ai-apps-being-used-to-make-deepfake-nudes

improvements in adversary capabilities, but they have not created entirely new threats. Malicious actors operated in all of the mediums, used all of the techniques, and worked towards all of the purposes discussed above prior to the advent of GenAI. Further changes are occurring over time, rather than being instantly adopted by all adversaries for all activities. As with legitimate business activities, adopting AI tools effectively takes time, and malicious cyber actors are no different.

"AI IS REPLACING THE MAJORITY OF ATTACK TECHNIQUES"

While AI tools have made certain malicious behavior, like disinformation campaigns, more efficient, they have not fundamentally changed the majority of attack techniques. The reality is that bad actors are still experimenting with AI capabilities. Contrary to the belief that AI instantly "supercharges" all bad actors, adoption remains measured, as many adversaries are still learning how to effectively use these tools. GenAI has lowered the bar, making basic techniques accessible to more people, particularly inexperienced "script kiddies,"[39] but it has not replaced traditional tactics, techniques, and methods.

"AI MAKES THE ADVERSARY SMARTER"

Some of the marketing hype around AI claims that it can make users "smarter." By extension, this increase should also apply to malicious actors. Yet, the way adversaries are using AI shows that AI is not making them smarter; it is making them more efficient. It is also lowering the bar for entry into malicious activity. Malicious actors may still struggle with language or coding accuracy, but AI enables faster, more accurate phishing, malicious website creation, and code generation.

"AI IS ALWAYS RIGHT"

AI suffers from a phenomenon called *hallucinations*, where it will respond to inputs with false information. This is due to a myriad of technical factors within the LLM architecture, but it essentially boils down to the AI's inability to reason and design limitations. LLMs such as ChatGPT are not designed with the ability to differentiate truth from fiction, instead prioritizing cohesive language output. For example, in 2023, an attorney was found to have used ChatGPT in their legal research, citing a case that did not exist.[40] This limitation would also impact adversaries ability to use GenAI reliably.

IMPROVING CYBERSECURITY: A PRACTICAL GUIDE

Although our analysis shows that GenAI enhances certain malicious tactics, it is ultimately just another tool that adversaries have added to their arsenal. Further, the most significant shift with GenAI is not in the sophistication of the tools themselves, but in the increased efficiency and reduced entry barrier they afford. As a result, defending against AI-enhanced threats may be harder than defending against standard threats, but it does not require revolutionary tools or techniques. This section offers guidance on actionable steps to strengthen cybersecurity against AI-enhanced threats.

MAINTAINING CYBERSECURITY FUNDAMENTALS

Despite advances in GenAI, implementing basic cybersecurity techniques still provides strong protection against malicious activity. Fundamental practices like multifactor authentication, regular software updates, offline data backups, strong

[39] Script kiddie is a demeaning term used to describe novice hackers who use existing scripts and software to carry out cyberattacks (https://us.norton.com/blog/emerging-threats/script-kiddie)

[40] https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c

passwords, and network monitoring still provide robust protection against cyber threats, including those incorporating AI. More sophisticated organizations can employ zero-trust architectures, endpoint detection, network traffic monitoring, and behavioral analytics. Therefore, maintaining or implementing these long-standing security measures while staying informed of AI developments is crucial.

AI-SPECIFIC DEFENSES

Beyond the fundamentals, employing some additional techniques against certain AI-enhanced cybersecurity threats will be necessary to manage cybersecurity risk effectively.

Guarding Against the New Generation of Phishing Emails

To counter the latest wave of AI-driven phishing, organizations should prioritize anti-phishing education that trains users to evaluate context more rigorously. For example, users should consider questions like: “Did I initiate this request?” or “Does this message align with my usual communications from this bank or retailer?” Encouraging critical thinking helps users spot red flags that go beyond mere typographical errors.[41] Given the accuracy of AI-generated phishing content, organizations should emphasize considering context over looking for superficial cues.

Cybersecurity vendors should support these educational initiatives, as advice about what to look for will change rapidly. Meanwhile, the industry can focus on understanding the underlying techniques and prompts of AI-generated phishing threats to enable more effective protection mechanisms. AI can play a role and evaluate emails by cross-referencing email content with known patterns of phishing behavior to assess the makeup of embedded links for potential inconsistencies.

Protecting Against Deepfakes

Technical Solutions

Defending against deepfakes is already difficult and will likely prove more challenging over time. A few technical tools leverage good AI to counter bad AI by deploying specialized models trained to detect deepfakes through advanced methods such as analyzing color histograms, Fourier transforms, and audio spectrums. However, while there are tools on the market that are promising, there is debate within the community about the efficacy of the tools, as well as their ability to integrate with vendors’ existing ecosystems. Absent some technological breakthroughs, organizations will likely have to rely on process-based solutions for the foreseeable future.

Multi-Channel Verification

Given the limitations of technical solutions, organizations should implement “old school” verification methods to prevent exploitation by deepfake technology. For example, in the case of financial transactions, arrangements above a certain amount should require dual verification, using a method different from the one used to initiate the request. Similarly, when dealing with suppliers, processes should mandate identity verification through more than two methods before altering payment details. For money-related requests made via phone or video call, using a pre-arranged authentication phrase known only to C-level executives and finance team members[42] or asking personal questions that only the actual person would know and that are not easily found online would provide an additional layer of protection.[43] This approach is particularly useful within companies, where verification through internal channels is

[41] https://www.mcafee.com/blogs/other-blogs/mcafee-labs/the-dark-side-of-gen-ai/

[42] https://blog.scilabs.mx/en/recommendations-for-preventing-audio-cloning-and-deepfake-fraud-in-corporate-environments/

[43] https://fortune.com/2024/07/27/ferrari-deepfake-attempt-scammer-security-question-ceo-benedetto-vigna-cybersecurity-ai/

possible. A verbal safe word can also be agreed upon to help verify identity.[44]

Heightened Awareness

Building heightened awareness among employees about deepfake manipulation is also a critical layer of defense. Organizations should encourage employees to adopt healthy skepticism and ask themselves questions like: "Is this call legitimate?" or "Am I being asked to do something unusual?" A key control is ensuring employees are familiar with their systems and processes. Many deepfake scams succeed because individuals fail to verify basic details. For instance, employees should confirm whether a call is expected, scheduled, or conducted on the usual platform. Additional signs of manipulation might include unexpected participants, changes in typical behavior (e.g. a colleague who normally engages in side chats suddenly staying silent), or inconsistencies in the meeting setup.

Beyond verifying system and process details, employees should evaluate the context and emotional tone of the communication. Questions like: "Does this video seem out of character?" or "Does this post make me feel unusually angry or compelled?" can prompt reflection and reduce impulsive responses. If in doubt, users should fact check the material to help confirm its legitimacy.[45] Additional layers of validation include reverse image searches, cryptographic signatures, and metadata verification such as through hashtags and blockchain digital fingerprints.[46]

However, as deepfake technology evolves, traditional indicators of manipulation, such as unnatural eye movement, misaligned features, or stiff facial expressions,[47] are becoming less reliable. To address this challenge, organizations should pair employee training with robust verification processes, ensuring employees are equipped to identify and respond to potential threats effectively.

The Pause, Take 9 campaign reinforces these principles by encouraging individuals to pause and assess situations before taking action, such as clicking a link, sharing sensitive information, or responding to unexpected results.[48] Simulated phishing tests, such as audio clones of executives, can enhance awareness and preparedness by mimicking real-world threats in a controlled environment.[49]

CONCLUSION

The rise of GenAI represents both opportunities and challenges in cybersecurity; empowering the community to leverage AI for innovation, efficiency, and enhanced defenses, while also enabling malicious actors to exploit the technology for a new dimension of AI-assisted threats. While GenAI lowers barriers to entry for adversaries and makes them more efficient, the foundational principles of cybersecurity remain integral to combating these threats effectively.

This JAR leverages the collective expertise of the CTA community to demystify the GenAI landscape, moving beyond the hype and providing evidence-based use cases. Through extensive collaboration, CTA members are actively addressing this critical topic across many collaborative efforts.

This report underscores that proactive measures,

[44] https://www.mcafee.com/ai/news/ai-voice-scam/

[45] https://defendcampaigns.org/ddcblog/aiandupcomingelections

[46] https://www.gendigital.com/blog/news/innovation/ai-elections-2024

[47] https://www.gendigital.com/blog/news/innovation/ai-elections-2024

[48] https://pausetake9.org/

[49] https://blog.scilabs.mx/en/recommendations-for-preventing-audio-cloning-and-deepfake-fraud-in-corporate-environments/

coupled with established best practices, are essential to mitigate the risks posed by AI-driven threats. Organizations must align their policies and practices to address the unique vulnerabilities associated with GenAI and can adopt a holistic approach that integrates technology, policy, and education.

The good news is that we have not seen anything truly new or revolutionary from the use of GenAI by malicious actors; GenAI has been used to enhance existing threats rather than create entirely new threats. As a result, organizations can continue to rely on enhanced versions of existing security practices to protect themselves. However, as malicious actors are just beginning to adopt GenAI tools, this window to prepare will not last forever. Organizations need to plan now for the changes that GenAI will make in the cybersecurity threat landscape.

DEFINITIONS

Unless otherwise cited, these definitions are derived from the Working Committee discussions held between August 22 and November 7, 2024. These definitions reflect the consensus of the Working Group.

- **Adversarial Threats:** Any threat against an AI model such that it tricks it to behave differently than the intended task. Such threats can be in the form of prompt injection attacks, evasion attacks, training data poisoning, reverse learning attacks. **NOTE** that adversarial threats are threats against the model. This is different from adversaries utilizing AI to generate malicious content (a.k.a. AI-Assisted Threats).

- **Agentic AI:** While Generative AI focuses on creating content, Agentic AI focuses on taking action. Agentic AI is a type of artificial intelligence that is connected to systems and Application Programming Interfaces (APIs) and can perform actions without human intervention. Agentic AI can use GenAI (e.g., the LLM aspect to write text) and therefore navigate complex scenarios and solve problems through reasoning.

- **AI Scams:** AI has reportedly been used in various types of scams for illicit financial gains. Not only have the generalized scams we've observed on a regular basis become more convincing, but the language used has also grown increasingly sophisticated. Additionally, deepfake technology has emerged as a powerful tool in these schemes.

- **AI-Assisted Threats:** This is when an AI model is used as a tool to assist malicious actors to generate content that is utilized for malicious purposes. Examples of AI-assisted threats include: malware generation, spam, phishing, voice clones, image diffusion, etc.

- **AI-Powered Bot Farm:** A group of automated accounts or bots that are coordinated by an AI system to perform activities in tandem.

- **Artificial Intelligence (AI):** AI is an interdisciplinary field of computer science that aims to design and train systems capable of performing tasks that would typically require human intelligence. To do this, AI relies on algorithms and computational models to process usually large sets of data to identify meaningful connections and patterns. It is defined by the Organization for Economic Co-operation and Development (OECD) as "a machine-based system that, based on explicit or implicit objectives, processes inputs to produce outputs like predictions, content recommendations, or decisions that can impact the physical or virtual environments."[50]

- **Audio Clone:** Synthetically generated audio that attempts to emulate a natural audio voice by

[50] https://one.oecd.org/document/GOV/SBO(2024)14/en/pdf

learning its characteristics. Such clones may or may not be used maliciously.

- **Deep Learning:** A subfield of ML that uses neural networks (“deep” meaning they contain many stacked layers forming high-level representation of unstructured data) to model complex patterns in data.

- **Deepfakes:** Synthetic media that have been digitally manipulated, including audio, images, and video. Deepfakes leverage techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content to imitate a real-world equivalent, often intended to deceive the intended audience into believing its authenticity.[51]

- **Fully Generated Images:** Images created entirely by a GenAI tool with no original photo as a base.

- **Generative Adversarial Network (GAN):** A GAN emanates in the category of Machine Learning (ML) frameworks, and uses deep neural networks to generate (after training) content that aims to preserve the likeness of the original data.[52]

- **Generative AI (GenAI):** A branch of AI that refers to systems capable of generating content such as text, music, images, and video. They differ from the more traditional Machine Learning (ML) where the goal is more discriminative such as for classification problems rather than a creative output.

- **Generative Pre-trained Transformer (GPT):** GPT is a model that involves pre-training on a large corpus of unlabeled text to learn general language patterns, followed by fine-tuning on specific tasks. It is trained in two stages: self-supervised language modeling to learn general patterns and supervised fine-tuning to adapt the model to specific tasks.[53]

- **Hallucinations:** The phenomenon in which LLMs provide incorrect information in a factual manner. A hallucination contains content that is at discord with factual knowledge sources.[54]

- **Inpainted Images:** These images begin with a real image in which GenAI is used to modify some aspects of the image. Detecting inpainting can be more challenging to detect as compared to fully generated images because its visibility largely depends on the extent and subtlety of the modifications made.

- **LLM (Large Language Model):** A type of AI that is based on deep learning to perform natural language tasks. Such models can be tuned to classify, generate, and translate language tasks. A key reason for their success stems from deriving context between tokens (words of the language) also known as *attention*.

- **Machine Learning (ML):** Classical machine learning (differentiated from other techniques such as Deep Learning)provides prediction and classification capabilities for specific tasks based on learning from historical data. Instead of programming specific instructions for each scenario, the computer uses models and algorithms to identify patterns in the data and learn from them.

- **Predictive AI:** The use of AI systems to analyze historical and real-time data to forecast future events or behaviors - e.g., to forecast spending or

[51] https://www.fsisac.com/hubfs/Knowledge/AI/FSISAC_Adversarial-AI-Framework-TaxonomyThreatLandscapeAndControlFrameworks.pdf

[52] https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/12/generative-artificial-intelligence-in-finance_37bb17c6/ac7149cc-en.pdf

[53] https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[54] https://www.fsisac.com/hubfs/Knowledge/AI/FSISAC_Adversarial-AI-Framework-TaxonomyThreatLandscapeAndControlFrameworks.pdf

anticipate fiscal risks.[55]

- **Prescriptive AI:** Goes a step further from predictive AI by not only forecasting outcomes but also suggesting courses of action to achieve desired goals or mitigate risks.[56]

[55] https://one.oecd.org/document/GOV/SBO(2024)14/en/pdf

[56] https://one.oecd.org/document/GOV/SBO(2024)14/en/pdf

CYBERSECURITY IN THE AGE OF GENERATIVE AI:

PART I — COMBATING GENAI ASSISTED CYBER THREATS

2025

CYBER THREAT ALLIANCE

POWERED BY THE CTA