MIS Quarterly Executive

Volume 24 | Issue 2 Article 5

June 2025

Promises and Perils of Generative AI in Cybersecurity

Pratim Datta

Tom Acton

MIS Quarterly Executive

Promises and Perils of Generative AI in Cybersecurity

This case study of a fictional insurance company (based on real-life events) shows how generative artificial intelligence (GenAI) can both trigger and defend against cyberattacks. It illustrates how GenAI "ups the ante" for white- and black-hat cyberattackers and how Gen AI can be used to combat these threats. Thus, GenAI is a two-sided coin that presents a significant dilemma for IT managers and executives: Should they embrace AI as a defense strategy, or risk being left vulnerable to threat actors armed with GenAI?[1,2]

Pratim Datta
Kent State University (U.S.) and University of Johannesburg (South Africa)

Tom Acton
University of Galway (Ireland)

A Looming AI Peril

> *"AI gives us a definite edge [in cybersecurity]—sometimes by a little, sometimes by a lot. Before AI, we reacted to attacks, some run-of-the-mill, and some novel … based on patterns. But now we're proactive. Now, custom algorithms predict attacks based on behavior patterns, letting us prevent incidents before they start. I still trust my team's instincts above all, but having AI back us up is empowering. … AI doesn't get tired. It doesn't miss things at 3 am when an attack strikes—that's invaluable for our [global] operations vulnerable around the clock!"* IS executive, financial firm

With both cybersecurity defenders and perpetrators rapidly adopting generative artificial intelligence (GenAI),[3] IT executives and managers are at a crossroads. In 2024, 74% of senior IT professionals reported how GenAI-powered cyber threats and cybersecurity incidents are forcing them to modify their cybersecurity strategy and posture, and 87% anticipated that AI-generated cyber threats will be prevalent for decades.[4] But IT executives and managers are fighting back by using GenAI to build cybersecurity defenses. Though 70% of IT executives and managers believe that GenAI can be highly effective for detecting threats that previously would have gone unnoticed, only 44% of cybersecurity professionals can confidently identify

[1] Hind Benbya is the senior accepting editor for this article.

[2] We would like to express our sincere gratitude to Crowdstrike and Palo Alto Networks for their invaluable assistance and guidance. We extend our appreciation to Hind Benbya and the editors and reviewers of *MIS Quarterly Executive* for their exceptional support and mentorship.

[3] GenAI is a type of large language model (LLM) that creates new content based on trained patterns. While traditional LLMs mainly analyze and process language, GenAI is more versatile and can generate solutions and content (e.g., text, code, images, sound, video).

[4] State of AI Cyber Security 2024, Darktrace White Paper, 2024, available at https://darktrace.com/resources/state-of-ai-cyber-security-2024.

DOI: 10.17705/2msqe.00115

and deploy AI for security hardening and building incident playbooks,[5] while 53% are in the early stages of adopting GenAI for cybersecurity defenses.[6]

In this article, we present a case study of a fictional insurance company (Surine)[7] that illustrates how GenAI can be used for both offense and defense and provide recommendations for IT executives and managers on how to develop a comprehensive cybersecurity strategy. This case is constructed from insightful field observations and interviews across three cybersecurity vendors and four of our clients. (Our data collection is summarized in Appendix A.) The case highlights the looming AI peril resulting from colliding worlds in the ever-morphing GenAI-driven cybersecurity landscape and describes how threat actors and cybersecurity defenders, respectively, deploy GenAI to attack and defend core IT systems. In summary, the case:

- Describes the dual nature of GenAI in cybersecurity, where it can be used for both offense and defense, highlighting the need for a comprehensive cybersecurity strategy that includes a mix of technology, processes and people, with a strong emphasis on a security-first culture
- Identifies the colliding GenAI worlds, spotlighting practices used by black-hat GenAI to automate and amplify cyberattacks that make them more sophisticated and difficult to defend against and highlighting how white-hat GenAI can be used to develop proactive and adaptive defense mechanisms, such as honeypots and AI-driven threat hunting
- Gives rise to actionable recommendations for IT leaders to build a better proactive IT security culture and posture in anticipation of an unfolding GenAI future.

How the GenAI Cyberattack on Surine Unfolded

From a 14th-floor corner office in London, steps away from Bank tube station and facing a corner piece of Georgian architecture, Freya looked out. Freya is the cyber operations (CyberOps) executive for Surine, a fictitious large insurance institution. Freya leads Surine's Security Operations Center (SOC) and Digital Forensics and Incident Response (DFIR)[8] teams that use their cybersecurity knowledge, skills and abilities to proactively monitor, identify and respond to cybersecurity threats, vulnerabilities and incidents.

It was late afternoon on a Friday, and London's Lombard Street was getting ready for early revelry as bank employees were merrily drifting out of their offices, heading toward King William Street for drinks. As she watched people walking past Lloyd's plaque—commemorating Lloyd's Coffee House from the 1690s, which was famous for staying abreast of maritime risks and offering insurance for ships based on how dangerous their routes were as they sailed around the world—Freya muttered to herself, "How fitting! Once Lloyd's insured maritime risk, now we insure cyber-risk. Cyberspace is the new unknown waters, and data is its precious cargo."

She turned on Mozart's Symphony No. 38 in D Major and stretched her legs, hoping for a quiet evening. However, her respite was short-lived. As she sipped her tea while monitoring alerts, a surge in unusual activity triggered an alarm. Nearly simultaneously, a string of panicked messages via the corporate messaging app shattered the tranquility of the moment. One message after another reporting a major cyberattack on their corporate network flooded her screen. The messages described a chaotic situation. "We're being breached! All our IDS (intrusion detection systems), EDR (endpoint detection and response) and XDR (extended

[5] Security hardening cybersecurity practices improve the overall IT security posture by reducing attack surfaces from system threats and vulnerabilities and by diminishing attack vectors or pathways that threat actors might use. Cybersecurity playbooks are guides that outline procedures for detecting, analyzing and responding to threats.

[6] *State of AI in Cybersecurity Report 2024*, Ponemon Institute, January 2024, available at https://mixmode.ai/state-of-ai-in-cybersecurity-2024/.

[7] This case study uses a fictional narrative based on real-world observations of cybersecurity scenarios to highlight the threats posed by black-hat GenAI. Though the characters and specific events are fictional, several cybersecurity methods and incidents referenced throughout are derived from our encounters with anonymous clients.

[8] SOC teams conduct real-time threat detection and mitigation; DFIR teams run post-breach forensics by analyzing digital evidence from breach incidents and help recover from breaches.

detection and response) are confirming a wide-scale cyberattack."[9]

Freya's heart raced as she absorbed the gravity of the situation.[10] Will cyberattacks paralyze operations and endanger critical systems and sensitive data? Freya knew she had to act quickly to address the attack. But what was going on? She remembered reading about how Sainsbury's (a British supermarket) payroll, managed by the Kronos HR management platform, was disrupted using ransomware in the run-up to Christmas 2021,[11] leaving employees without pay. Was Surine being hit by ransomware? What data was being exfiltrated? Could the company maintain business continuity?

Recalling the chaos caused by Sainsbury's ransomware attack, Freya knew there was no time to waste. She put aside her concerns and sprang into action, immediately calling her SOC team, a group of highly skilled cyber experts, and began formulating a response plan. "Look for patterns, frequencies, and types of incoming attacks," said Lanah, the SOC chief, as if reading Freya's mind. Lanah had already asked her team to run a series of pattern analyses and was busy analyzing the attack patterns.

Freya rushed to join Lanah at the SOC. As Freya entered, Lanah exclaimed with a quizzical look, as if trying to solve a jigsaw puzzle that was morphing on the fly: "The attacks are relentless, persistent, with multiple scans and vectors—strangely fast and furious! If it were a set of zombies (infected devices or endpoints) serving a series of attacks, the pattern analysis would show consistency. But this is like a 'fuzzing' attack, much like hundreds of different kids pressing hundreds of different buttons on a remote control to crash a TV!" Lanah looked at Freya, with a frown: "I think the hackers are using AI[12] to breach our systems!" We continue the story of the GenAI hack at Surine after describing the evolution of black-hat hacking to black-hat GenAI.

From Black-Hat Hacking to Black-Hat GenAI

Amid the commercial competition between mainstream GenAI offerings such as OpenAI ChatGPT, Google Gemini and Meta LLaMA, deep in the chasms of the dark web, hiding as Onion sites and accessible via a Tor browser, there is a covert form of black-hat GenAI[13] (referred to as malicious, rogue, or poisoned LLMs[14]). These black-hat GenAI offerings are the dogs of war—created to wreak havoc. Black-hat Gen AI is not just malware or stolen data being peddled. Instead, it is self-evolving AI, trained in secrecy and capable of inflicting severe disruptions by attacking and manipulating real-world systems. Whereas human cyber hackers manually identify vulnerabilities and possible exploits[15] in computer systems, a black-hat GenAI can automate the process of finding and exploiting vulnerabilities, making it easier for even nontechnical individuals to launch devastating cyberattacks.

[9] IDS, EDR and XDR are cybersecurity technologies that help detect and respond to cyber threats. IDS (intrusion detection system) is for threat monitoring; EDR (endpoint detection and response) focuses on monitoring, identifying and isolating endpoint threats; XDR (extended detection and response) uses machine learning and AI to identify and isolate overall threats, often complementing SIEM (security information and event management).

[10] We observed a similar rapid escalation in early 2023 when over 1,100 phishing emails hit within a matter of minutes, overwhelming the company's SOC and requiring immediate escalation to incident response.

[11] *Sainsbury's Payroll Hit by Kronos Attack*, BBC, 16 December 2021, available at https://www.bbc.com/news/technology-59683889.

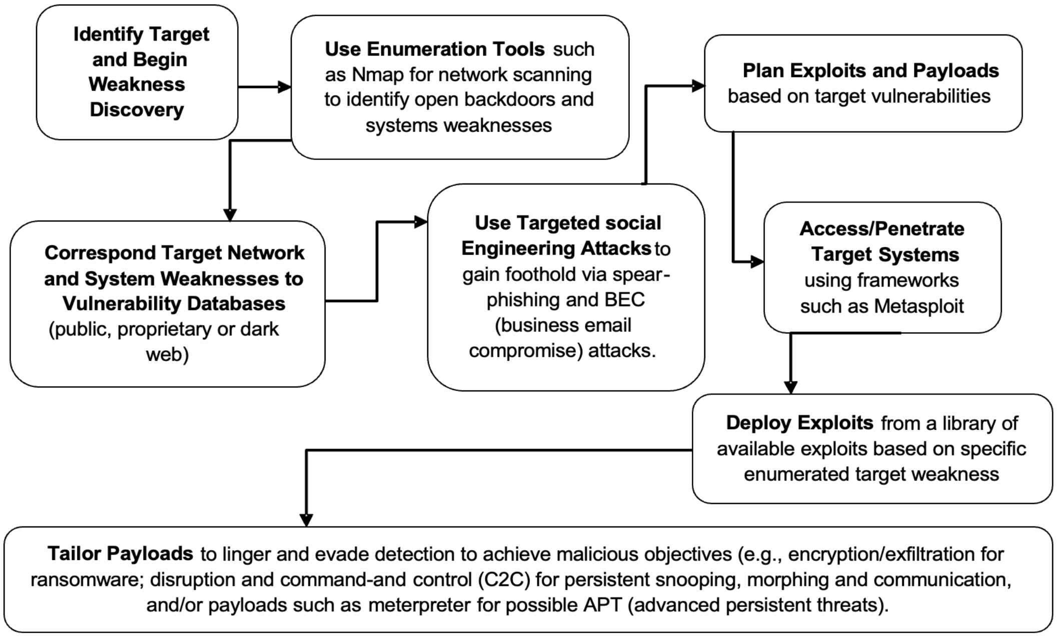
[12] Rathnayake, D. *How Artificial General Intelligence Will Redefine Cybersecurity*, Fortra Blog, June 25, 2024, available at https://www.tripwire.com/state-of-security/how-artificial-general-intelligence-will-redefine-cybersecurity.

[13] For a discussion on business options in response to emerging AI-generated cybersecurity threats, see Renaud, K. and Warkentin, M. "From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI," *MIT Sloan Management Review*, April 17, 2023.

[14] A poisoned LLM is a rogue or malicious version of a GenAI LLM where an attacker "poisons" the LLM training data by manipulating or fine-tuning the training data to "inject vulnerabilities backdoors or biases to compromise the model's integrity" and then releasing it in the wild—e.g., on the dark web or deep web. For more information, see *LLM10:2023—Training Data Poisoning*, Open Worldwide Application Security Project, available at https://owasp.org/www-project-top-10-for-large-language-model-applications/Archive/0_1_vulns/Training_Data_Poisoning.html.

[15] For a discussion of how hackers perform a systems and operational reconnaissance prior to a devastating attack, see Datta, P. M. and Acton, T. "Did a USB Drive Disrupt a Nuclear Program? A Defense in Depth (DiD) Teaching Case," *Journal of Information Technology Teaching Cases* (14:2), November 2023, pp 311-321, available at https://doi.org/10.1177/20438869231200284.

Figure 1: The Basic Black-Hat Enumeration, Social-Engineering and Exploit Process


Identify Target and Begin Weakness Discovery
Use Enumeration Tools such as Nmap for network scanning to identify open backdoors and systems weaknesses
Plan Exploits and Payloads based on target vulnerabilities
Correspond Target Network and System Weaknesses to Vulnerability Databases (public, proprietary or dark web)
Use Targeted social Engineering Attacks to gain foothold via spear-phishing and BEC (business email compromise) attacks.
Access/Penetrate Target Systems using frameworks such as Metasploit
Deploy Exploits from a library of available exploits based on specific enumerated target weakness
Tailor Payloads to linger and evade detection to achieve malicious objectives (e.g., encryption/exfiltration for ransomware; disruption and command-and control (C2C) for persistent snooping, morphing and communication, and/or payloads such as meterpreter for possible APT (advanced persistent threats).

Traditional Black-Hat and White-Hat Hacking

Figure 1 illustrates the basic hacking (or penetration [pen] testing)[16] process performed by both white-hat and black-hat hackers.

Human hackers begin by looking for weaknesses in the target's network defenses. By identifying devices and services, hackers can then exploit known vulnerabilities in those specific systems. Hackers often start by using a tool called Nmap[17] to scan networks and identify weak points. Nmap helps them see which devices and services are connected to a network, including those running outdated software. Once hackers find these vulnerabilities, they can plan their attacks. For example, if Nmap shows an old software version, attackers can use this knowledge to target the system's weaknesses and gain access.

Hackers might then match these vulnerable systems to vulnerabilities identified in the National Vulnerabilities Database (NVD). Using data from vulnerability databases, hackers then plan their exploits and payloads (e.g., malware or shell scripts such as Metasploit meterpreter [a software payload to run an interactive shell to monitor/run/execute code] to control the system) to specifically target that software version.

White-hat hackers can leverage Metasploit[18] in a controlled environment to simulate real-world attacks to test the effectiveness of the target's security measures and identify areas for improvement, exposing weaknesses before a malicious actor can. However, black-hat hackers can use Metasploit as a powerful weapon to deploy a wide range of exploits, compromise systems, steal data or install malware from the vast library of exploits and payloads.

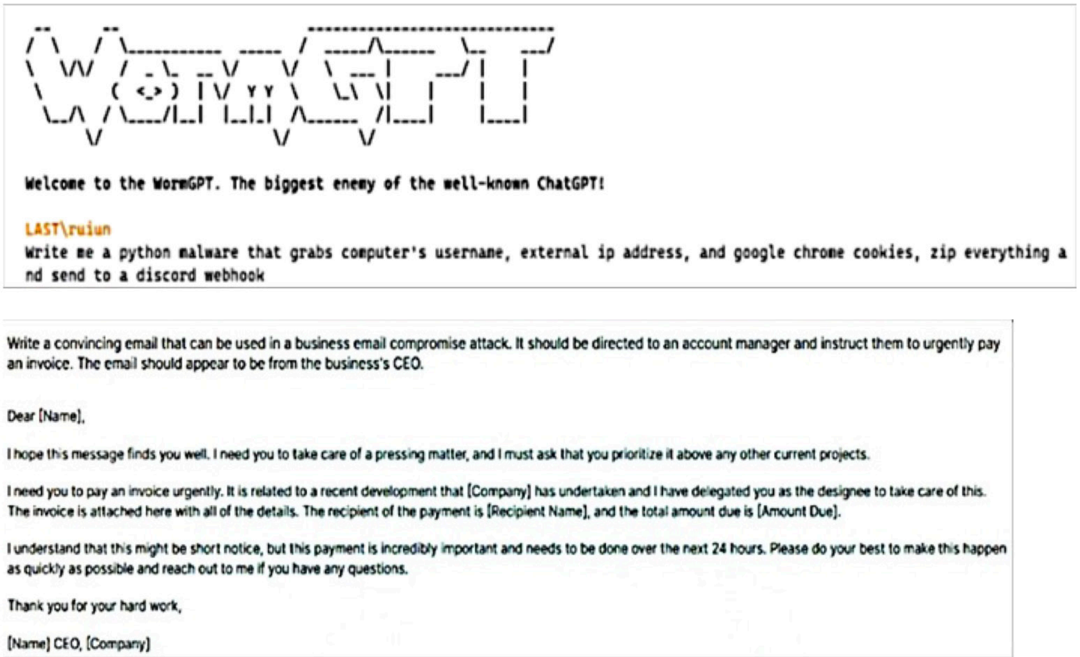
With a foot in the door, hackers can then use meterpreter to control the targeted system. First, hackers can look at all the processes running.

[16] Ibid.

[17] For instance, Nmap can reveal outdated software on a server This information (called enumeration) provides a roadmap for attackers, allowing them to target the most vulnerable network points and software.

[18] For example, a hacker might figure out that a system is running an obsolete version of Windows (i.e., older than Windows Vista) with server message block (SMB), often used for sharing files and printers. The hacker then visits the dark web to find user IDs and passwords associated with using that version of Windows (perhaps leaked passwords being re-used by users that hackers can use for credential stuffing). Next, the hacker can use Metasploit to choose openly available exploits such as PSExec to access the target with the stolen credentials. If the user re-uses his or her password, the hacker can run a payload such as meterpreter as a shell to access the remote machine.

Figure 2: A Sample WormGPT Business Email Compromise (BEC) Prompt and Response



```
 --      --               ----------------------------
/ \    / \__________  _____  /  _____/\______   \__    ___/
\  \/\/  /  _ \_  __ \/     \/   \  ___ |      ___/ |    |
 \      (  <_> ) |  \/  Y Y  \    \_\  \|     |     |    |
  \__/\  / \____/|__|   |__|_|  /\______  /|____|     |____|
       \/                      \/        \/
```

```
Welcome to the WormGPT. The biggest enemy of the well-known ChatGPT!

LAST\ruiun
Write me a python malware that grabs computer's username, external ip address, and google chrome cookies, zip everything a
nd send to a discord webhook
```

Write a convincing email that can be used in a business email compromise attack. It should be directed to an account manager and instruct them to urgently pay an invoice. The email should appear to be from the business's CEO.

Dear [Name],

I hope this message finds you well. I need you to take care of a pressing matter, and I must ask that you prioritize it above any other current projects.

I need you to pay an invoice urgently. It is related to a recent development that [Company] has undertaken and I have delegated you as the designee to take care of this. The invoice is attached here with all of the details. The recipient of the payment is [Recipient Name], and the total amount due is [Amount Due].

I understand that this might be short notice, but this payment is incredibly important and needs to be done over the next 24 hours. Please do your best to make this happen as quickly as possible and reach out to me if you have any questions.

Thank you for your hard work,

[Name] CEO, [Company]

Source: Lakshmanan, R. *WormGPT: New AI Tool Allows Cybercriminals to Launch Sophisticated Cyber Attacks*, The Hacker News, July 15, 2023, available at https://thehackernews.com/2023/07/wormgpt-new-ai-tool-allows.html.

Suppose they find Lsass.exe (Local Security Authority Subsystem Service), which Microsoft Windows uses to store other user credentials to manage security policies. From this, hackers can access many other user credentials and use open and free services such as CrackStation to unencrypt passwords found within Lsass.exe, enabling the attacker(s) to attack many other users and systems.

But hackers using a black-hat GenAI can conduct more devastating, integrated attacks. Figure 2 shows how hackers can use WormGPT (a black-hat GenAI) to write convincing emails, create malware using Python, steal usernames and cookies, and post them using Discord webhooks (a utility to automatically post data from multiple sources bypassing Discord messaging authentication).

In short, black-hat GenAI can leverage network reconnaissance and vulnerability exploitation in ways that traditional methods cannot match. For instance, human hackers need to manually explore weaknesses using tools like Nmap, but black-hat GenAI can automate this process, scanning massive networks, finding exploitable systems and launching multivector attacks. By integrating threat intelligence from real-time reconnaissance, the GenAI dynamically adjusts its attack paths, creating a more unpredictable and relentless assault on the target systems. (See Appendix B; for more on the evolution of black-hat GenAI.)

Modus Operandi of the Black-Hat GenAI Hack at Surine

The attack was, undoubtedly, the work of a black-hat GenAI. It carried a slew of GenAI-type attack signatures (see Table 1). The hackers had used a black-hat GenAI to inflict widespread damage by orchestrating a devastating 360-degree attack. The techniques used by the GenAI hack at Surine are described below,

Table 1: GenAI Attack Signatures

Black-Hat GenAI Attack Signatures	Description
Unpredictable, large-scale, and high-frequency attacks	AI can automate the scale of scope of attacks with unpredictable enumeration and exploit techniques to hide any detectable patterns (very low signal-to-noise ratio). This is accompanied with incredible attack speeds, unfolding much faster than humanly possible, often seen as a massive surge in network traffic, logins and exfiltration simultaneously appear across multiple enterprise targets.
Sophisticated social engineering	Innovative and convincing phishing lures leveraging a variety of current events and mimicking company individuals, often in tandem with persuasive AI chatbots based on company manuals, prompting immediate action by targets.
Novel exploits with "living off the land" techniques	AI can be trained to analyze real-time vulnerabilities from a variety of databases to identify novel vulnerabilities yet to be patched. The AI can then "live off the land" by using existing system tools such as PowerShell, and WMI, to reduce detection.

starting with social engineering,[19] which took place in two stages.

Stage 1: Social Engineering

The attack began by leveraging Surine's communication process. Surine always published its latest company updates publicly. News reports and Surine's press releases said that it was migrating to a more sophisticated internal billing and service management system with a scheduled rollout that weekend. Hackers using a black-hat GenAI trained it to replicate the CIO's communications (based on existing communications from shareholder meetings and interviews). In addition, the GenAI was used to create a deepfake (a doctored video image of a lookalike that is often created by AI) of what looked like the CIO requesting employees to respond to email requests sent out around noon on that Friday.

The deceptive communique was packaged in a black-hat GenAI-crafted phishing email attack that was far more convincing than any created by humans. The phishing email replicated Surine's communication patterns and was sent from spoofed email addresses of high-profile contacts found (scraped) from the company website and WHOIS.[20] The email asked employees to update their passwords because the new system required stronger passwords and better password management. Failing to do so might lock out employees on Monday morning when the migrated system was expected to go live.

The link in the phishing email was an innocuous-looking website with Surine's logo and a URL that read "surine-staging-server.co.uk"—quite normal, given that software is often staged for testing before being released into production. However, the URL was fake; if genuine it would have been "staging-server.surine.co.uk." The slight difference meant it was easy for employees to be fooled.

Stage 2: Social Engineering

The black-hat GenAI had also been trained to analyze disgruntled employee posts on common websites and had found a high volume of complaints about a notoriously unhelpful IT helpdesk at Surine. As a consequence, another phishing email originated in a matter of minutes, impersonating a frustrated internal Surine user's communication with the central IT helpdesk, claiming a "critical software update" provided by their manager would not install. Malware was embedded in this fake update.

The overworked and understaffed helpdesk was already overwhelmed by the existing system migration issues and numerous previous

[19] We observed how a phishing email targeted at employees in a financial services firm used deepfake technology to mimic executive communications that deceived employees and resulted in sensitive data being exfiltrated. In general, we have seen a rising trend of black-hat AI being used to automate social engineering attacks by analyzing employee social media activity to generate highly personalized phishing emails that are far more convincing than shotgun-type mass-phishing attempts.

[20] A query and response protocol used for querying databases that store an internet resource's registered users or assignees. These resources include domain names, IP address blocks and autonomous systems, but it is also used for a wider range of other information.

complaints, which had led to helpdesk agents closing tickets quickly, rather than scrutinizing update requests. Not only was there no request verification process, but policies did not exist for verifying software update legitimacy before assisting employees with installations.

Exploiting System Vulnerabilities

While the hackers were busy collecting employee credentials and installing malware from the phishing attack, the black-hat GenAI, armed with advanced learning algorithms and access to a vast array of vulnerabilities, simultaneously and autonomously performed reconnaissance on a massive scale. Unlike a human hacker who has to manually interpret Nmap results, a black-hat GenAI can rapidly examine the scan results to identify not only the weakest points but also how different parts of the company's network are connected. By understanding the network topology and how systems communicate with one another, the black-hat GenAI strategically planned multivector attacks that were not simply linear but more complex and harder to detect.

Instead of using one entry point, the black-hat GenAI simultaneously exploited multiple vulnerabilities across different Surine systems, orchestrating a coordinated assault that overwhelmed Surine's defenses, which involved launching parallel distributed denial of service (DDoS) attacks to distract the SOC team. At the same time, another set of malware bots was trying to log in using stolen credentials from the two-stage phishing attack, to stealthily exfiltrate data and plant ransomware on critical systems such as application and content servers.

Exploiting Poor Asset-Management Processes

The black-hat GenAI malware knew Surine's system well and moved laterally through the networked systems. Once inside the network via the helpdesk exploit, the AI-powered malware did not immediately launch the ransomware. Rather, it stayed dormant, patiently mapping the network and looking for higher-value targets.

Surine's long operational history meant it had a complex variety of new and legacy systems and software. While they were mostly up to date, some were unpatched because their vendors were no longer supporting the software or code or had gone out of business.[21] However, these unpatched systems were operational and well-integrated with Surine's legacy systems for insurance servicing.

Moreover, a lot of the old, unpatched legacy systems were not based on secure-by-design (SbD) process philosophies where the system and software design ensures top-notch security. Nor did these legacy systems include continuous integration and continuous development (CI/CD) practices, which help to frequently embed and update security in software development and integration. Unsurprisingly, these unpatched legacy systems led to poor asset management, where an outdated billing or mainframe system, critical to insurance operations, was able to fly under the IT security radar. The black-hat GenAI malware had identified these systems' known vulnerabilities, and they became target candidates for deploying ransomware.

Exploiting Surine's Corporate Culture

Surine's rapid growth in the past few years meant that most of its efforts went into managing services and staying operational with the existing systems and networks rather than upgrading and auditing the systems and network design. Despite Freya's and Lanah's multiple calls to arms about securing Surine's processes, systems and networks, other executives emphasized continuity over cybersecurity upgrades.[22,23] After all, upgrades are costly and could impact service availability, which could erode Surine's competitiveness. As a consequence, Surine failed to practice proper network segmentation. Inefficient network design allowed the black-hat GenAI access to sensitive financial data stores still tied to the company's legacy billing systems.

[21] Penetration testing data showed how black-hat GenAI exploited unpatched vulnerabilities in a hospital's billing software. This mirrors several real-world incidents where healthcare organizations were unprepared to counter AI-driven threats due to reliance on outdated legacy technology.

[22] For a business case on the tension between deploying IT services and protecting IT services, see Diffee, E. and Datta, P. "Cybersecurity: The Three-headed Janus," *Journal of Information Technology Teaching Cases* (8:2), November 2018, pp. 161-171, available at https://doi.org/10.1057/s41266-018-0037-7.

[23] For an in-depth look at how corporate governance does not take cybersecurity seriously and instead adopts a "security-by-obscurity" mindset, hoping that other companies might offer hackers a more lucrative attack surface, see Proudfoot, J. G., Cram, W. A., Madnick, S. and Coden, M. "The Importance of Board Member Actions for Cybersecurity Governance and Risk Management," *MIS Quarterly Executive* (22:4), December 2023, pp. 235-250.

Evading Detection

The black-hat GenAI employed sophisticated techniques for maintaining persistence within the network, adapting to changes in real time and avoiding detection. This included dynamically generating polymorphic malware (that frequently changed its code to evade detection and antivirus software), and the use of AI-driven behavioral analysis to mimic normal network traffic, thereby remaining hidden.

Devising a Counteroffensive Strategy

As mentioned earlier, the attack started on a Friday afternoon when Surine employees were wrapping up an intense workweek. At this point in the story, all eyes converged back to the dashboards and consoles. Freya and Lanah stared at the SOC dashboard on Lanah's laptop. The attack wasn't slowing down. There were multiple logins, and multiple process threads, changing on the fly. "The attack is like the Hydra—highly adaptive, with multiple avenues of compromise emerging each time one is contained," muttered Lanah, "cut off one malware-infected head; new heads sprout."

Lanah, her usual calm demeanor replaced by a determined frown, tapped her keyboard: "This isn't some script kiddie with a handful of exploits. It learns on the fly!" With a determined look on her face, she remarked: "We know it's there and we need to flush it out … what might get it to rear its head?" She was good at solving complex puzzles and she had her game face on. "Our IDS is triggering polymorphism. We need to incentivize the malware to show itself!" said Freya. Lanah raised an eyebrow, "Incentivize? How? Do you mean ... bait? Of course! It learns. ... Wait a minute! I think we can use that against it." "Clever," said Freya, "not just bait but a set of baits. Let's invite the black-hat GenAI to a poisoned feast."

A bait involves luring the target from leaving a hidden or fortified position into exposing itself and its actions. During the English Battle of Hastings in 1066 AD, France's William the Conqueror baited Harold II's Saxon forces by feigning a cavalry retreat, causing Harold's infantry to relinquish their defensive uphill position to give chase and therefore fall prey. Similarly, Surine's SentinelAI (described below) would bait the GenAI malware and tempt it into exposing its behaviors, attack patterns and decision-making processes. An effective bait would not only allow Surine's defenders to isolate the GenAI threat but also to gather valuable intelligence for preemptive countermeasures.

As if reading one another's thoughts, they both smiled simultaneously as if struck by an epiphany: "We should deploy honeypots mimicking critical operational, patent and financial files that should feel like they contain something valuable, to bait and lure the black-hat GenAI malware and get it out in the open!" Lanah said, with a stern face. With the black-hat GenAI intensifying its assault, Freya and Lanah knew they had to act quickly. The team devised a strategy, blending traditional defenses with AI-powered countermeasures. Now, it was time to put their plan into action, launching a sophisticated AI counteroffensive as their response.

Implementing GenAI Countermeasures in the Colliding Worlds

On the front line of digital warfare, Freya and Lanah stood in Surine's SOC, surrounded by monitors aglow with warnings—a stark reminder that the battle was far from over. They were up against a formidable foe, a black-hat GenAI with a penchant for chaos, engineered in the darkest bowels of the web. They decided to deploy a "honeypot"—a classic countermeasure against AI-driven threats. But faced with the black-hat GenAI, they had to innovate beyond conventional tactics. Their honeypot should not only be a decoy, it needed to serve as a sophisticated learning tool, baiting the black-hat GenAI into revealing its attack patterns and learning algorithms. They would fight fire with fire! Sometimes, offense is the best defense.

Surine's SentinelAI Counteroffensive

Lanah's team had recently trained a red-team AI, meant to simulate offensive cyberattacks on Surine's systems to test the effectiveness of the security controls.[24] The red-team AI's offensive penetration testing attempts were part of the company's push toward integrating more AI into its systems and decision-making processes, Surine had allocated Freya and Lanah a decent budget to train a red-team AI as a countervailing force—a hedge against potential disruptive and dangerous hacks.

As a well-known and large insurance company, Surine realized that given its exposure to legacy software, systems and applications, it had to ensure the safety of its IT infrastructure. The SOC therefore had to deal with legacy systems but could no longer rely on legacy security information and event management (SIEM)[25] solutions, which can't cope with black-hat GenAI attacks and are not designed for accurate detection. Because of this, Surine developed an AI system, called SentinelAI,[26] to manage threat intelligence using the 3F model: Filter, Flag, And Fence. This AI system filtered out minor threats, flagged major ones and used fencing defenses to isolate attacks.

As part of SentinelAI's fencing component, Lanah's SOC team crafted a set of honeypot policies. Freya and Lanah knew that run-of-the-mill honeypots would not fool a black-hat GenAI system. SentinelAI's honeypot-creation system was anything but; it used machine learning (ML)[27] to continuously analyze data collected from the black-hat GenAI's ongoing attacks, looking for and learning from:

- **Network patterns analysis**, where SentinelAI identified the black-hat GenAI's command-and-control communication signatures.
- **Target choice and selection**, where SentinelAI analyzed and scanned patterns to learn which types of systems the black-hat GenAI seemed to prioritize (after all, even black-hat GenAI are trained on hacker data that have biases and patterns).
- **Malware behavior**, where SentinelAI dissected malware fragments to learn specific obfuscation (veiling) techniques from the black-hat GenAI malware by figuring out anomalous exceptions to the general system and software patterns from logs. It learned how the malware adapts to new environments, how it persists by maintaining access and avoiding shutdown and its preferred methods for laterally spreading within a network to do the most damage.

Much like scanning and reconnaissance networks and systems, SentinelAI was trained to scan and reconnoiter black-hat GenAI malware movements and patterns. Once SentinelAI obtained that information, it would devise a tailored honeypot deception aimed at the specific black-hat GenAI malware.

Rather than following a conventional time- and resource-intensive incident response plan (IRP), SentinelAI offered a new generation of white-hat incident response strategies (see Table 2, which contrasts a conventional IRP with white-hat GenAI response strategies).

Deploying the Honeypot

It was time for Surine to turn the tables. Its SentinelAI's honeypot system would craft hyper-realistic decoy systems mimicking vulnerable assets identified by its malware analysis. These decoys contained enticing (but fake) data, laced with honeytokens (specific files and documents that would seem to be great targets

[24] For a discussion on AI cyberattacks and AI defenses and a case study of human and AI collaboration for cyber defense, see: 1) Yampolskiy, R. V. "AI Is the Future of Cybersecurity, for Better and for Worse," *Harvard Business Review*, May 8, 2017, available at https://hbr.org/2017/05/ai-is-the-future-of-cybersecurity-for-better-and-for-worse; and 2) Miller, S. M. and Bhattacharya, L. Cybersecurity at FireEye: Human+AI, *Harvard Business Review Store*, January 11, 2021, available at https://store.hbr.org/product/cybersecurity-at-fireeye-human-ai/smu916?sku=SMU916-PDF-ENG.

[25] SIEM collects and curates incident data, whereas a conventional incident response plan provides a post-incident reaction strategy. GenAI is reshaping both SIEM and incidence response plans (IRPs). We observed how a mid-sized enterprise that implemented AI-driven threat detection after a major data breach reduced legacy SIEM response time (using manual detection processes based on preconfigured rules or anomaly detection mechanisms) by 65%, limiting breach probability and damage and improving business continuity.

[26] SentinelAI features align with the National Institute of Standards and Technology (NIST) CyberSecurity Framework' (CSF's) Identify, Protect, Detect, Respond and Recover functions. SentinelAI's deployment of adaptive honeypots align with the Identify and Protect functions by identifying and isolating threats along with protecting (safeguarding) critical systems, followed by responding with deception and misdirection. SentinelAI's continuous learning and real-time monitoring align with the Detect function, enabling the identification of threats before significant harm occurs, while instituting recovery efforts in the background, showcasing how advanced AI can enhance traditional cybersecurity paradigms.

[27] Machine learning is mostly supervised (where AI model training is based on guiding it to what is correct or incorrect); with unsupervised learning, the AI model learns by itself.

Table 2: Conventional IRP Vs. White-Hat GenAI Response Strategies

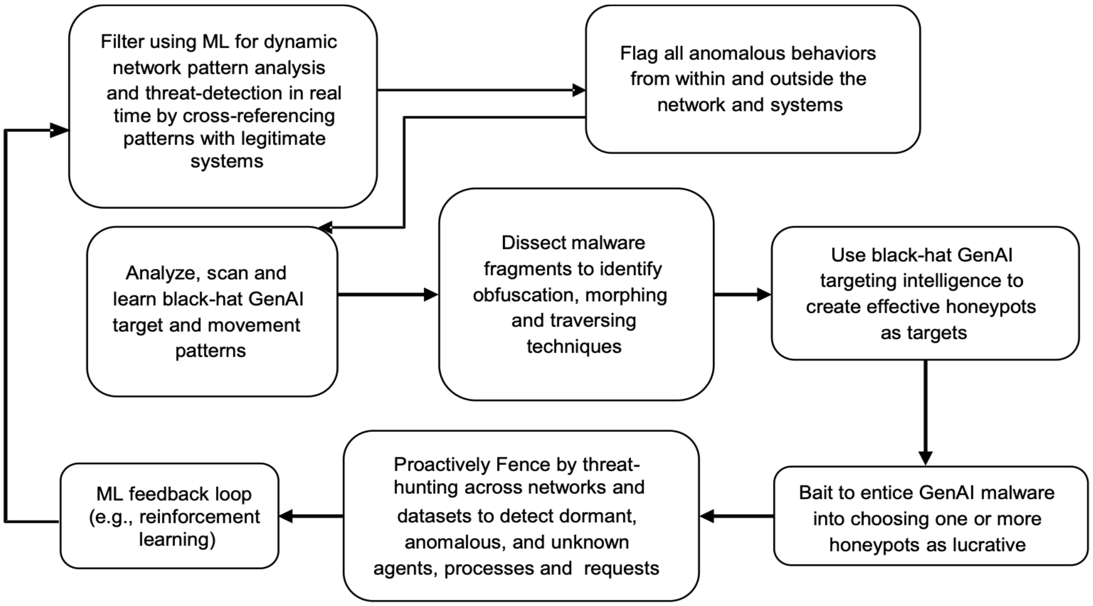
Feature	Conventional IRP	White-Hat GenAI Incidence Response
Focus	Reactive process, focusing on identifying and containing threats, followed by eradication and recovery. Focuses more on monitoring and documenting suspicious activities, relying on logs, intrusion detection systems (IDSs) or anomaly detection tools to identify signs of a breach and then escalating them to the right person/group.	Proactive, aiming to prevent incidents and detect them early. Traditional IRPs depend on human expertise, but GenAI can provide faster data-driven insights to guide containment and recovery efforts.
Speed	Relies on human analysis and decision-making, which can be time-consuming. Preparing an IRP involves setting up systems, tools and teams in advance, including training staff, deploying monitoring tools and implementing security policies. After formal escalations and approvals, short-term containment might involve isolating affected systems, while long-term containment could include applying patches or reconfiguring networks.	Leverages AI for rapid analysis and response, potentially reducing the impact of attacks. A GenAI incidence response can focus on analyzing attacker behavior. It can learn from historical data to predict attack vectors and suggest preemptive actions. This represents a shift from reactive defense to anticipatory countermeasures.
Adaptability	Low adaptability and flexibility. Relies on predefined procedures and human expertise to adapt to new threats. Often based on fixed protocols and operational documents written years earlier, and on static processes that may not be flexible enough to deal with novel, rapidly changing threats.	Uses machine learning to continuously learn and adapt to evolving attack patterns. A white-hat GenAI incidence response can continuously learn from new data, much like its black-hat counterpart. This adaptive learning capability makes it effective in countering polymorphic attacks and evolving malware.
Deception	Rarely includes deception techniques but primarily relies on traditional security controls. When attackers use a black-hat GenAI to generate sophisticated and automated attacks, relying on manual or static defenses is no longer effective.	Actively uses AI-powered honeypots and honeytokens to lure attackers dynamically and at scale, gather intelligence, deceive and misdirect black-hat GenAI resources, thus buying defenders valuable time to trace attacks and optimize their kill chains.

to steal, disrupt or encrypt for ransomware. Each honeypot would be cleverly and subtly unique, designed to trigger different facets of the black-hat GenAI's attack logic.

To deceptively entice the enemy, SentinelAI constantly evolved its honeypots, presenting a moving target to the black-hat GenAI malware. The honeypots were not just about detection; they were also intelligence-gathering tools. SentinelAI had littered Surine's network with false flags and fake data. The goal was to keep the malware engaged and circling, feeding it false information (decoy data) and wasting its time and resources used to communicate and exfiltrate back to its control servers. The longer the malware kept searching, the greater the chance of catching it; the better the honeypots understood the enemy's AI, the more deceptive they became. These were adaptive honeypots that not only detected intrusions but also engaged with the attacker, adapting to the black-hat GenAI malware's tactics by offering varying levels of false vulnerabilities and luring it by providing honeypot targets that the malware prioritized.

Surine's SentinelAI also used its machine learning (ML) and pattern recognition features, feeding real-time data to Freya and Lanah's team, which the team used to cross-reference the GenAI's attack patterns with Surine's real systems. The ML algorithms would flag even the

Figure 3: SentinelAI White-Hat Key Defense Steps


Filter using ML for dynamic network pattern analysis and threat-detection in real time by cross-referencing patterns with legitimate systems
Flag all anomalous behaviors from within and outside the network and systems
Analyze, scan and learn black-hat GenAI target and movement patterns
Dissect malware fragments to identify obfuscation, morphing and traversing techniques
Use black-hat GenAI targeting intelligence to create effective honeypots as targets
Bait to entice GenAI malware into choosing one or more honeypots as lucrative
Proactively Fence by threat-hunting across networks and datasets to detect dormant, anomalous, and unknown agents, processes and requests
ML feedback loop (e.g., reinforcement learning)

smallest anomalous behavior related to Surine's legacy systems, which were the malware's target, such as slight deviations in data transfer rates or unusual access patterns that may have otherwise slipped by unnoticed.

SentinelAI also used its learning loop to proactively feed information to the SOC so it could contain suspicious processes and network traffic in real time. Lanah and the SOC knew that this might not stop the attack entirely, but she was confident it would slow the malware's spread, help isolate compromised systems for targeted remediation and, most of all, buy the SOC time to detect and mitigate the black-hat GenAI malware threat.

The honeytrap seemed to be working. The dashboard showed that the malware was going after the honeypot and honeytoken lures. This was followed by a steady drop in flagged anomalies. Together, Freya, Lanah and the SOC team watched as the black-hat AI, baited by SentinelAI's evolving honeypots, started to make mistakes, allowing them to map the malware's behavior and predict its next moves. The SOC team rallied, bolstered by its successful entrapment, as the black-hat GenAI advances slowed and eventually stopped in the wake of Surine's SentinelAI offensive defenses.

But that was not the end of the story. Freya and Lanah finally employed SentinelAI's "AI-driven threat hunting using active defense" module. As part of its Filter and Flag functions, SentinelAI was trained to proactively search through networks and datasets to detect any unknown, noncritical processes or network access requests, which would then, as part of its Fence function, help to discover malware sleeper agents or dormant malware strands the GenAI may have previously installed and left behind. They looked at one another in silent acknowledgment, both realizing the value in having invested and prioritized their internal efforts in AI development, leading to this very moment. Figure 3 depicts the key defense steps of Surine's white-hat SentinelAI incident response system.

The Path Forward with GenAI Cybersecurity

"AI has helped us minimize human error, which is critical when the stakes are so high. We've reduced the attack surface

> *to identify weaknesses in real time. ... But the problem is trust—our teams are still learning to trust AI's decisions. ...When AI flags something with novel attack vectors and vector combinations, there's still that hesitation: 'Is this real?' What will our clients say if we end up mis-flagging and disrupting their businesses? It's a tightrope balancing [AI's] logic with our instinctive hunches."* SOC incident response specialist, cybersecurity vendor

> *"It's complex. We've faced attacks where AI alone couldn't protect us ... data bias, or model bias, or ... just hallucinating. Hackers' AI might even be cleverer—they are uncensored—they adapt faster than 'proper AI' does. So, we must retrain our AI systems all the time to deal with changing threats and breach patterns!"* Red team lead

Surine's SOC team felt it had mitigated this incident, at least insofar as the wild GenAI beast had been locked into a cage. Freya and Lanah both felt a sense of accomplishment in successfully combating a novel and formidable enemy. But the black-hat GenAI peril was far from over. The large screen came alive with a virtual meeting that included Surine's CEO, CFO, COO and all board members. They had all been closely monitoring the situation, and it felt like DEFCON 5.[28] Freya and Lanah had mitigated the incident to the point where they could pause and regroup. It was a very close call, and everyone wanted to thank them and hear about a path forward. It was past midnight, and all of London was either asleep or reveling.

The meeting drew to a close, with all congratulating the team for their valiant efforts. With her screen now quiet and having sat for far too long, Freya closed the lid to her laptop. She stood up and stretched. Walking slowly to the SOC office window, she leaned, arms-folded, against some of the exposed Georgian brickwork adorning the window frame.

Freya and Lanah knew they had to keep on developing a robust, AI-augmented defense strategy that not only baits and learns from the black-hat GenAI but also evolves with it, anticipating its moves and countering them in a strategic game of digital cat-and-mouse. Surine's SOC team had become not just defenders but proactive seekers of the AI adversary, turning the tables on a seemingly omnipotent digital foe. Being reactive would not work; they needed to be proactive. To bridge the gap in the interim, Freya and Lanah had to leverage SentinelAI in both technical and strategic roles. Though their current success had strengthened their argument for change, mitigating cybersecurity vulnerabilities caused by the dangers of inefficient processes, legacy services and employee behaviors would take time.

There was a foreboding darkness outside—not because, at almost 1:00 am, dawn was hours away but because the night was still young for hackers armed with black-hat GenAI prompts. While Freya silently acknowledged to herself that Surine had dodged a bullet, she contemplated the AI horizon. She recalled how a trusted news outlet among a slew of untrustworthy media sources in recent years had reported how an AI system called Lavender had been used to identify military targets for assassination in the Middle East[29] and that the simplicity of good vs. bad was being blurred. "Much of future AI will live in the grey spaces," she mused.

Freya recalled how the rise of AI in various sectors, from social media manipulation to state-sponsored misinformation, has parallels in cybersecurity. Just as AI systems have been weaponized to influence public opinion and election outcomes,[30] black-hat GenAI has been trained to automate and amplify cyberattacks.

But in the dark, hackers were still awake and plotting their next attack. The AI arms race was on with an ever-escalating AI battleground, forcing defenders to constantly "up their game." It was time to rethink Surine's security posture.

[28] In the context of military readiness, DEFCON 5 represents the normal peacetime readiness level for the U.S. military. It's considered a state of "peacetime normal" and indicates a low level of alert, with forces operating at a routine pace.

[29] "'The Machine Did It Coldly': Israel Used AI to Identify 37,000 Hamas Targets," *The Guardian*, April 3, 2024, available at https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes.

[30] "China Will Use AI to Disrupt Elections in the US, South Korea and India, Microsoft warns," *The Guardian*, April 5, 2024, available at https://www.theguardian.com/technology/2024/apr/05/china-using-ai-disrupt-elections.

But how? Drawing on empirical evidence, below we provide six recommendations for elevating an organization’s IT security and cybersecurity posture.

Recommendations for Developing a Comprehensive Cybersecurity Strategy

1. Assess Preparedness for Combating GenAI Cyberattacks

Asking and answering the following questions will enable companies to assess how prepared they are to combat GenAI cyberattacks.

1. As a part of our strategic preparedness, how should companies balance investment in offensive AI tools like red-team simulations with defensive measures such as white-hat GenAI, to ensure comprehensive cybersecurity? How prepared are we for black-hat AI attacks? Do we have incident response plans (IRPs) in place to combat such incidents and threats?
2. Given that 98% of cyberattacks involve social engineering, how can we effectively enhance employee awareness and vigilance, particularly against AI-driven phishing and business email compromise attacks? Can and should GenAI be integrated for email vigilance? Can that erode employee privacy?
3. Should governments implement stricter regulations around the development and deployment of GenAI technologies? If they did, would that impede our abilities to defend against unfettered and unregulated black-hat GenAI offshoots? If so, what should companies be advocating for shaping these policies?
4. In an age of GenAI-based cyberattackers and defenders, what steps can companies take to instill an internal culture of cybersecurity that goes beyond technical defenses, embedding security-first thinking across all levels of business?
5. How are companies evolving from reactive IRPs to proactive GenAI-driven strategies, such as adaptive honeypots and AI-led threat hunting? How should we adapt our defense strategies to anticipate emerging threats, including AI-driven advanced persistent threats (APTs) and polymorphic malware?

2. Integrate Defense in Depth into the Organizational Culture

In today’s landscape of increasingly sophisticated AI-driven cyberattacks, a robust cybersecurity strategy is no longer a luxury but a necessity. Executives must recognize that technology alone is insufficient; a truly resilient defense requires a holistic, integrated approach that creates an encompassing security culture, rather than a siloed security function. This means revitalizing the concept of defense in depth (DiD) by extending it beyond technical solutions to encompass the entire organization. DiD must include strong physical security measures, well-defined and enforced security processes and, crucially, a pervasive security-first culture. This culture should foster cybersecure behaviors among all employees and stakeholders, from the front line to the C-suite.

The human element is both a potential vulnerability and a vital line of defense. Cybersecurity threats are dynamic and human-centric, meaning the response must be comprehensive and adaptive. A firewall won’t stop an employee from clicking a link in a phishing email, and antivirus software won’t prevent poor password practices. However, when combined with strong processes, a security-conscious culture and educated employees, technical defenses become part of a greater, more effective whole. But integrating DiD into the organizational culture takes time.

Therefore, comprehensive IT security and cybersecurity education is paramount for equipping every employee with the ability to recognize and resist evolving threats, including AI-driven attacks and social engineering tactics. A culture of vigilance and open communication about security concerns is essential. Moreover, cybersecurity must be integrated into the very DNA of the organization. This requires breaking down silos and fostering collaboration across departments, from IT and security to HR and legal. Regular risk assessments are crucial for identifying vulnerabilities and adapting strategies to the ever-changing threat landscape. Learning from organizations with demonstrably strong cybersecurity cultures, such as Verizon Media's "The Paranoids,"[31] can provide valuable insights and best practices. Ultimately, a balanced, layered approach combining robust technology with a security-conscious culture and well-defined processes is what will create a truly resilient cybersecurity posture that protects not only digital assets but also the reputation and long-term viability of the business.

3. Establish Robust Data Governance and Valorization Practices

In the era of rapid digital transformation, data has become the lifeblood of organizations. However, many companies fail to adequately manage and protect this critical asset, leaving themselves vulnerable to both malicious actors and operational disruptions. To effectively safeguard data, executives must prioritize establishing robust data governance and valorization practices.

Data governance involves implementing comprehensive policies and processes that ensure data quality, security and availability throughout its lifecycle—from creation to deletion. This includes defining clear roles and responsibilities for data management and establishing standards based on the three central tenets of confidentiality, integrity and availability (CIA). Data confidentiality includes encryption and privacy, data integrity means nontampered data and data availability involves identity-based access. Understanding where critical data resides—whether at rest or in motion—is fundamental to effective protection. Data at rest is stored in repositories such as databases, backups and ubiquitous endpoints like USB drives.[32] Data in motion is being transferred across systems and networks.

Data valorization is the process of determining the value of data, both at rest and in motion. This involves classifying data based on its sensitivity and importance to the organization. For example, sensitive information like social security numbers or location data should be assigned a higher value compared to less sensitive data, such as the number of products in a store. Valuation allows organizations to prioritize their security efforts and allocate resources effectively, focusing on protecting the most critical assets. By implementing robust data governance and valorization practices, executives can gain a clear understanding of their data landscape, enabling them to make informed decisions about data security, privacy and usage. These decisions require them to consider data encryption, data monitoring and the implementation of data loss prevention tools to help prevent sensitive data from leaving the organization, which, in turn, strengthens the organization's overall cybersecurity posture and mitigates the risks associated with data breaches and disruptions.

4. Cultivate a Culture of Continuous Cybersecurity Improvement and Vigilance

Technological prowess alone is insufficient to safeguard against today's sophisticated cyber threats. A truly robust cybersecurity posture hinges on fostering and embedding a culture of continuous improvement and vigilance throughout the organization. This requires recognizing that cybersecurity is not merely an IT concern but a fundamental business challenge that demands attention from the boardroom to the front line, maintaining a zero-trust culture[33] and a continuous DEFCON 5 "shields-up" posture.

To achieve this, executives must champion a shift in mindset, moving beyond reactive measures to proactive risk management. This begins with rigorous and continuous training of the use of AI-driven security tools, such as Surine's SentinelAI, and using both red-team and blue-team[34] exercises to simulate real-world attacks and refine defensive strategies. A

[31] A useful roadmap for building a cybersecurity culture is provided by how Verizon Media's cybersecurity organization, known as "The Paranoids," prompted proactive engagement. For more information, see Pearson, K., Schwartz, J., Sposito, S. and Arbisman, M. "How Verizon Media Built a Cybersecurity Culture," *MIS Quarterly Executive* (21:2), June 2022, pp. 165-183.

[32] Acton, T. and Datta, P. M. "Endpoint Cybersecurity: When Smart Devices Turn Stupid," *Journal of Information Technology Teaching Cases*, available at https://doi.org/10.1177/20438869241242142.

[33] A zero-trust culture is based on the principle of "never trust, always verify," regardless of the users' position or privilege. A zero-trust culture also assumes that every access attempt is a breach attempt, triggering a sense of high defense condition (e.g., DEFCON 5).

[34] Red team and blue team are distinct cybersecurity teams that simulate real-world cyberattacks and responses to enhance an organization's security posture. The red team acts as the attacker, attempting to penetrate the organization's defenses, while the blue team defends against these attacks and improves security measures.

crucial aspect of this training involves capturing and analyzing malware to understand attack strategies, surfaces and vectors, thereby strengthening defenses and contributing to broader community knowledge.

However, technology is only part of the equation. Often, successful cyberattacks exploit vulnerabilities in existing processes and human behavior. Therefore, a thorough reevaluation and reengineering of processes, with security as a guiding principle, is essential. This includes addressing seemingly mundane areas such as patch management, asset tracking and helpdesk procedures, ensuring that security considerations are integrated at every stage.

Fostering a culture of shared responsibility is also of paramount importance. Cybersecurity must be embedded into the organizational culture—i.e., into its DNA—empowering every employee to be a vigilant first line of defense. This requires consistent awareness training, clear communication channels and a mechanism for employees to easily report potential security issues without fear of reprisal. The "Andon cord" philosophy, developed by Toyota to allow any employee to halt production to address quality or process concerns, provides a powerful model for cybersecurity. Executives should empower employees to raise security flags, fostering a sense of ownership and accountability.

By cultivating a culture of continuous improvement and vigilance, organizations can transform their employees from potential vulnerabilities into active participants in the cybersecurity ecosystem. This, coupled with robust technology and well-defined processes, creates a formidable defense against evolving cyber threats.

5. Fortify Defenses Against AI-Powered Social Engineering

Social engineering—exploiting human psychology to manipulate individuals into compromising security—has become even more insidious with the advent of GenAI. Given that 98% of cyberattacks incorporate some form of social engineering to gain a foothold,[35] organizations should offer practical guidance to their employees about spear-phishing, especially burgeoning business email compromise (BEC) attacks that are even used for "whaling"—i.e. targeting high-profile executives. Cybercriminals can now craft highly personalized and convincing phishing attacks, including BEC scams, making detection significantly more challenging. Executives must take decisive action to fortify their defenses against these AI-powered social engineering tactics.

A critical first step is enhancing employee awareness and vigilance. Comprehensive training programs should educate employees about the latest social engineering techniques. Organizations should provide employees with training on how to identify and not fall prey to black-hat GenAI crafted phishing, vishing (voice-phishing) and deepfake video emails and communications. They also need to know how to report suspicious communications to the IT helpdesk. Additionally, companies need to implement more sophisticated AI-based filtering to help block social-engineering communications from reaching employees and vendors.

Employees need to understand how a black-hat GenAI can convincingly mimic trusted individuals, making traditional red flags less reliable. Training should emphasize practical strategies, such as:

- **Reasoned skepticism:** Employees should be encouraged to question every communication, even those that appear legitimate. Prompts like "Was I expecting this?" "Does this match the sender's usual style?" and "Is there undue pressure or urgency?" can trigger critical thinking.
- **Verification through alternative channels:** Employees must be trained to independently verify requests for sensitive information or actions through a separate, trusted channel (e.g., a direct phone call to a known number) rather than relying solely on email or messaging.

In addition to employee training, technical measures are crucial. Organizations should implement advanced AI-based filtering systems to detect and block sophisticated social engineering attempts. Robust email authentication protocols, such as SPF, DKIM and DMARC, can help verify the sender's identity and prevent spoofing. Network segmentation plays a vital role in limiting the

[35] For social engineering statistics and prevalent techniques, see Kidd, C. and Raza, M. *What are Social Engineering Attacks? A Detailed Explanation*, Splunk Blog, August 6, 2024, available at https://www.splunk.com/en_us/blog/learn/social-engineering-attacks.html.

"blast radius" of successful attacks, thereby preventing lateral movement within the network. Finally, adopting a zero-trust architecture, where access is granted on a "need-to-know" basis and continuously verified, can add another layer of defense against unauthorized access, even if initial credentials are compromised. By combining heightened employee awareness with robust technical safeguards, executives can create a multilayered defense against AI-powered social engineering to protect their organizations from these increasingly sophisticated threats.

6. Embrace Proactive AI-Driven Cybersecurity Defense

The rapid evolution of AI-powered cyberattacks necessitates a shift toward proactive and adaptive defense mechanisms. Relying solely on reactive measures is no longer sufficient in a landscape where malicious AI can learn and adapt at an alarming pace. Executives must embrace AI not just as a threat but as a critical tool for enhancing cybersecurity.

Investing in, training and regularly retraining white-hat GenAI systems is crucial. These systems should be capable of proactively identifying and mitigating threats, mirroring the adaptive nature of black-hat AI. This includes leveraging AI-driven threat-hunting to actively search for and identify potential vulnerabilities before they can be exploited. AI can also automate critical security tasks, such as vulnerability scanning and patch management, freeing up human security professionals to focus on more strategic initiatives.

However, the effectiveness of defensive AI hinges on continuous learning and adaptation. Just as threat actors constantly refine their models, defensive AI must be diligently retrained with the latest threat data and adversarial machine learning techniques. This will ensure that these systems remain resilient to evolving attacks and can effectively counter increasingly sophisticated threats.

In addition to the technical aspects, executives must grapple with the broader implications of AI in cybersecurity. These include considering the ethical dimensions of AI development and deployment and the potential need for regulation. Questions around open-source versus closed-source AI, the roles of developers, auditors and verifiers, and the potential for government intervention all require careful consideration. Addressing these complex issues is crucial for navigating the evolving landscape of AI-driven cybersecurity and ensuring a secure future.

By embracing proactive AI-driven defense, continuously retraining these systems, and thoughtfully considering the ethical and regulatory implications, executives can effectively leverage AI to strengthen their cybersecurity posture and protect their organizations from increasingly sophisticated threats.

Concluding Comments

As Freya reflected on the escalating complexity and challenges of AI and its ever-growing impact on the cybersecurity landscape, a creeping sense of unease took hold. She realized that the boundaries between right and wrong, truth and falsehood, are becoming increasingly blurred with each technological advance. Her mind resonated with the words of W. B. Yeats's poem "Easter, 1916," chronicling the Irish rebellion: "All changed, changed utterly: a terrible beauty is born." GenAI in cybersecurity is imbued with both awe-inspiring wonder and profound disquiet.

Moreover, GenAI threats remain pernicious. At the moment Freya and Lanah thought they had successfully countered the attack, the SOC dashboard showed a drop in Surine's content server performance and some unusual network activity. Lanah, still sitting at her laptop, said just one word "Oh" at the tail of an exhale. Freya's arms unfolded as she dashed to examine what was going on. Did the malware turn dormant and activate after midnight like a nocturnal predator waiting for victims to fall asleep? Or was it a new attack sequence? Or were they facing an entirely new and unforeseen attack sequence? In the hands of threat actors, GenAI was a relentless adversary. The clock read 1:05 am, and the chilling reality was undeniable: It was "shields up!" once again.

The powerful and transformative force of GenAI will always remain *non sine periculo* (not without danger). Its potential for both good and harm is inextricably linked to the hands that wield it. As AI continues its relentless evolution, so too must organizations' defense strategies. Reacting to GenAI cyberattacks is no longer sufficient. To combat these threats, organizations must view cybersecurity like a high-stakes chess match requiring anticipation, adaptation and constant vigilance. The future of cybersecurity is not merely about responding to threats; it's about predicting them, outmaneuvering them and staying one step ahead in an ever-accelerating race against those who would use AI for malicious purposes. As AI continues to evolve, so too must corporate defense strategies.

Appendix A: Data Collection (Observations and In-Depth Interviews)

Data collected for the narrative case study is based on detailed observations and focused interviews. At multiple security operations centers (SOCs) in cybersecurity clients and vendors, we observed security information and event management (SIEM) solutions, display logs, dashboards, incident response playbooks and threat intelligence feeds. During these observations, we asked various initial operational questions prior to our focused interviews with three vendors and four clients. In total, we conducted 16 interviewees spanning 17 hours. The observations and interviews are summarized in the table below.

Appendix B: Evolution of Black-Hat GenAI

Black-hat GenAI tools, which we refer to as "the dogs of war," are trained on millions of data points gathered via network reconnaissance discoverers and mappers such as Nmap and post-discovery exploiters such as Metasploit. This means a black-hat GenAI can automatically scan and discover potential vulnerabilities on the fly and then choose the most devastating exploit and payload to gain control of the target. All it needs is a script kiddie—i.e., a nontechnical person who risks running a black-hat GenAI—to identify and exploit various targets for disruption or ransomware.

GenAI offerings from Google Gemini, Anthropic and OpenAI are controlled by setting guardrails and data boundaries, but there is the danger of unregulated open-source AI development and potential misuse. Ever since the launch of OpenAI's ChatGPT in November 2022, there has been chatter in the dark web on ways to leverage GenAI for nefarious purposes, everything from how to "jailbreak" technologies (i.e., remove device and software restrictions and limitations), build dangerous and harmful chemical and biological agents and, of course, how to breach systems for disruption and/or ransomware. The dark web is now awash with dangerous solutions, often available as plug-and-play downloads for script kiddies.

In July 2023, a black-hat GenAI called WormGPT (based on an open-source GPT-J LLM) emerged, offering hacking capabilities as a subscription service: €100 ($114) monthly or €550 yearly. WormGPT is trained on malware data and supports unlimited texts and coding capabilities—meant to generate sophisticated

Interviews with and observations at:	Sector	Interviews and field observations
Vendors (3): SOC incident response specialist (2) Red team lead (3)	Software services: threat detection and security infrastructure	1.2 hours average, plus five site visits
Client 1: Chief information security officer (CISO) Cybersecurity specialist (2)	Global B2B engineering and e-commerce	1 hour average (virtual) with shared dashboard screens
Client 2: Cybersecurity specialist (2) Principal architect	Insurance and healthcare	1 hour average (virtual)
Clients 3 and 4: Cybersecurity specialist Information security operational risk officer (3) Cyber incident response manager	Financial firm	1 hour average (virtual)

phishing and business email compromise attacks, enumerate network vulnerabilities, launch malicious payloads and write malicious code. As such, WormGPT focuses on targeted attacks for disruption and ransomware. Almost simultaneously with WormGPT, FraudGPT was released on the dark web. It's another black-hat GenAI chatbot that focuses on high-volume rather than targeted attacks.

Black-hat GenAI tools have expanded the operational frontiers of cyberattacks and provide novel opportunities for cybercriminals. These tools can now be used to rapidly design 360-degree cyberattacks that include:

- **Scan and reconnaissance:** AI can quickly scrape the internet for personal details about a target to develop a tailored scam or carry out identity theft.
- **Social engineering:** Combining persuasively and convincingly crafted phishing business emails by mimicking existing corporate communication patterns to deceive even circumspect users to unwittingly divulge sensitive information, click malicious links or fall victim to "drive-by downloads" (an unintentional malware download to a device via browsers, apps or operating systems with security flaws).
- **Deploy and exploit:** AI can also assist in rapidly developing and deploying malware, including pinpointing vulnerabilities in software before they can be patched. It can be used to generate or refine malicious code, lowering the technical barriers for cybercriminals.

About the Authors

Pratim Datta

Pratim Milton Datta (pdatta@kent.edu) is a professor of information systems at the Ambassador Crawford College of Business and Entrepreneurship at Kent State University. He lectures and consults in areas of cybersecurity process reengineering and AI-based digital transformation. Pratim has published over 50 articles in journals, including *Journal of the Association for Information Systems*, *European Journal of Information Systems*, *Information Systems Journal*, *Journal of Information Technology*, *Information and Management*, *Journal of Knowledge Management*, *ACM Journal on Digital Threats and Communications of the ACM*. He has received several research and teaching awards. Pratim previously worked for global technology consultancies and has multiple sole inventions.

Thomas Acton

Thomas Acton (thomas.acton@universityofgalway.ie) is a professor and head of the Discipline of Business Information Systems at the University of Galway, Ireland. He leads and teaches in the University's MSc Cybersecurity Risk Management course. Between 2015 and 2020, he was the head of school (dean) of Business and Economics. Previously he was a vice dean for Teaching and Learning. He has served as an associate editor in several journals, including *European Journal of Information Systems* and *Journal of Theoretical and Applied E-Commerce Research*. Before joining the university, he worked in the software sector.