MITSloan
Management Review

From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI

It's time to replace traditional, rule-based approaches to cybersecurity with "smarter" technology and training.

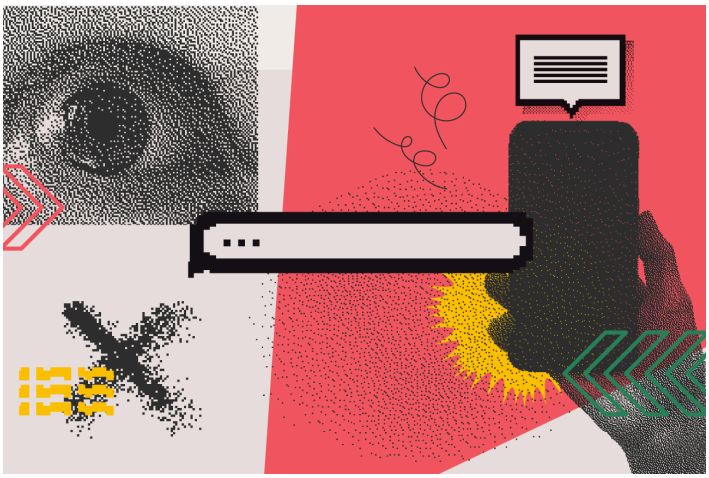
Karen Renaud
Merrill Warkentin
George Westerman

• Reprint #64428 • sloanreview.mit.edu

From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI

Karen Renaud, Merrill Warkentin, and George Westerman

It's time to replace traditional, rule-based approaches to cybersecurity with "smarter" technology and training.



For the past several years, cybercriminals have been using artificial intelligence to hack into corporate systems and disrupt business operations. But powerful new generative AI tools such as ChatGPT present business leaders with a new set of challenges.

Consider these entirely plausible scenarios:

- A hacker uses ChatGPT to generate a personalized spear-phishing message based on your company's marketing materials and phishing messages that have been successful in the past. It succeeds in fooling people who have been well trained in email awareness, because it doesn't look like the messages they've been trained to detect.
- An AI bot calls an accounts payable employee and speaks using a (deepfake) voice that sounds like the boss's. After exchanging some pleasantries, the "boss" asks the employee to transfer thousands of dollars to an account to "pay an invoice." The employee knows they shouldn't do this, but the boss is allowed to ask for exceptions, aren't they?
- Hackers use AI to realistically "poison" the information in a system, creating a valuable stock portfolio that they can cash out before the deceit is discovered.
- In a very convincing fake email exchange created using generative AI, a company's top executives appear to be discussing how to cover up a financial shortfall. The "leaked" message spreads wildly with the help of an army of social media bots, leading to a plunge in the company's stock price and permanent reputational damage.

These scenarios might sound all too familiar to those who have been paying attention to stories of deepfakes wreaking havoc on social media or painful breaches in corporate IT systems. But the nature of the new threats is in a different, scarier category because the underlying technology has become "smarter."

Until now, most attacks have used relatively unsophisticated high-volume approaches. Imagine a horde of zombies — millions of persistent but brainless threats that succeed only when one or two happen upon a weak spot in a defensive barrier. In contrast, the most sophisticated threats — the major thefts and frauds we sometimes hear about in the press — have been lower-volume attacks that typically require actual human involvement to succeed. They are more like cat burglars, systematically examining every element of a building and its alarm systems until they can devise a way to sneak past the safeguards. Or they're like con artists, who can build a backstory and spin lies so convincingly that even smart people are persuaded to give them money.

Now imagine that the zombies become smarter. Powered by generative AI, each one becomes a cat burglar, able to understand the design of your security systems and devise a way around them. Or imagine a con artist using generative AI to engage interactively with one of your employees, build trust, and dupe them into falling for the con.

This new age of AI-powered malware means companies can no longer use best-practice approaches that may have been effective mere months ago. Defense in depth — the strategy of installing the right security policies, implementing the best technical tools for prevention and detection, and conducting awareness drives to ensure that staff members know the security rules — will no longer be enough. A new era has dawned.

Using combinations of text, voice, graphics, and video, generative AI will unleash unknown and unknowable innovations in hacking. Successful defenses against these threats cannot yet be automated. Your company will need to move from acquiring tools and establishing rule-based approaches to developing a strategy that adapts to next-level AI-generated threats in real time. This will require both smarter tech and smarter employees.

Match Generative AI Against Generative AI

Companies should use generative AI both to strengthen their defensive capabilities and to accelerate their ability to respond to new threats in real time.

First, consider a company's perimeter defenses. Businesses already use malware databases to detect new threats. These databases of malware signatures are constantly refreshed by vendors but are not tailored to each company's unique situation. Hackers formerly used one-size-fits-all exploits (ways into a company's systems); now they will use AI to tailor their exploits to an individual company's vulnerabilities. Deceptive emails will push on the right pressure points; they will also be highly believable because the AI-driven language in new phishing attacks will use public information to tailor each message to the target company. So rather than dismissing an email immediately as fraudulent, even savvy employees are more likely to see enough specifics to be convinced that the message is legitimate.

OpenAI (the maker of ChatGPT) and others are releasing tools such as GPTZero and ZeroGPT that will enable companies to detect whether newly generated text (with no identifiable signature) has been produced by generative AI. By integrating these tools into mail servers, companies can improve the likelihood of blocking next-level automated phishing messages. These tools should be tailored to each company's needs and be frequently fine-tuned to maintain vigilance, just as an intelligent human gatekeeper needs to stay up to date on the latest threats. Over time, security tool vendors will incorporate these technologies, but in the meantime, many companies run the risk of being hacked and suffering major financial and reputational losses. Therefore, it's important to consider making internal changes in the short term rather than waiting for off-the-shelf solutions in the market to catch up.

Second, real-time detection has to improve, fast. Many companies rely on pattern detection to repel attacks. The problem is that the patterns are informed by attacks that have already happened. To counteract attacks driven by generative AI, a new era of "smart" prevention is needed.

Generative AI has the potential to improve companies' ability to rapidly detect anomalies in behaviors or actions, by employees or anywhere within company systems. Employee behaviors — evidenced by the systems they log into, how much data they access, and with whom they communicate

via email — tend to be predictable from day to day, matching their job tasks. This can be thought of as their behavioral footprint. If the footprint changes suddenly without their job description changing, it could signal, for example, a potential ongoing hacking attempt or insider misbehavior. Using generative AI in combination with other AI tools, companies can then identify the extent of the damage — or determine that no breach has occurred. New extensions and third-party applications built on the GPT-4 foundation have already been announced, so expect new AI-based security tools to be available soon.

Train People for Smarter Attacks

Security-aware human behavior remains critical to cyber safety, but people continue to make mistakes. Many awareness campaigns describe existing threats and provide a set of rules to follow: Don't click on links, be sure to use strong passwords, and patch all software, to name just a few. Numerous annual industry surveys have identified employees as the weakest link. Call center employees, for instance, can be fooled by people who have just enough information or provide the right type of emotional pitch. According to one estimate, as many as 82% of breaches involve human behavior.

In the era of generative AI, awareness training needs to shift from policies that mandate behavior to knowledge-based preparedness that allows employees to detect new threats. That is, employees need to know enough about hacking to graduate from rule followers to active defenders. With the advent of generative AI tools, traditional information security policies have outlived their usefulness; a rule-based approach is no longer appropriate.

Truck drivers cannot maintain safety merely by following traffic laws; they also need to adapt to road conditions, by allowing more distance during rainy conditions or slowing down in the fog, for example. Likewise, employees must apply situational knowledge to resist new cybersecurity challenges from generative AI. This requires that companies go beyond training people on what to do and what not to do. They also need to help them understand how to stay secure in a very challenging new world.

Companies need to move away from a compliance-based strategy to one in which new employee skills are developed. This may require instructor-led training, either in person or online, to develop knowledge that surpasses traditional rule-based policies. The current one-size-fits-all training, even when followed by a quiz, is passive. The AI era requires training that attunes employees to real or potential scenarios and includes live discussions on how to respond.

Beyond playing defense through employee awareness training, companies also need to follow Sun Tzu's wisdom that "defense is the planning of an attack." Imagine the worst case; think the unthinkable. Use AI-based models to hypothesize potential threat vectors or triggers to watch for. One strategy: Form a SWAT team of your best IT people, or even experts from other companies, to brainstorm how bad actors could potentially penetrate your defenses. Then find ways to improve your technical capabilities and employee awareness around such events. One survey found that companies are transitioning from the traditional red team/blue team (offense/defense) cyber war-gaming approach and instead adopting a more collaborative "purple team" approach to build deeper understanding of emerging attack methods and learn what works and what doesn't.

The best defense against AI-powered hacks is likely to be AI-informed. This means not just faster and more robust defense strategies, but genuinely smarter strategies for your technology and your people. You won't beat the smart zombies by making your fences higher. But you can augment your traditional defenses with new AI-powered tools, and you can rethink your traditional defense methods and rule-based training. In the new and scary world of HackGPT, you should get started now to keep your company safe and secure.

About the Authors

Karen Renaud is a computing scientist at the University of Strathclyde in Glasgow, working on all aspects of human-centered security and privacy. Merrill Warkentin, an ACM Distinguished Scientist, is a W.L. Giles Distinguished

Professor and the Rouse Endowed Professor of Information Systems in the College of Business at Mississippi State University. George Westerman is a senior lecturer at the MIT Sloan School of Management and founder of the Global Opportunity Initiative in MIT's Office of Open Learning.

MITSloan Management Review

PDFs ■ Reprints ■ Permission to Copy ■ Back Issues