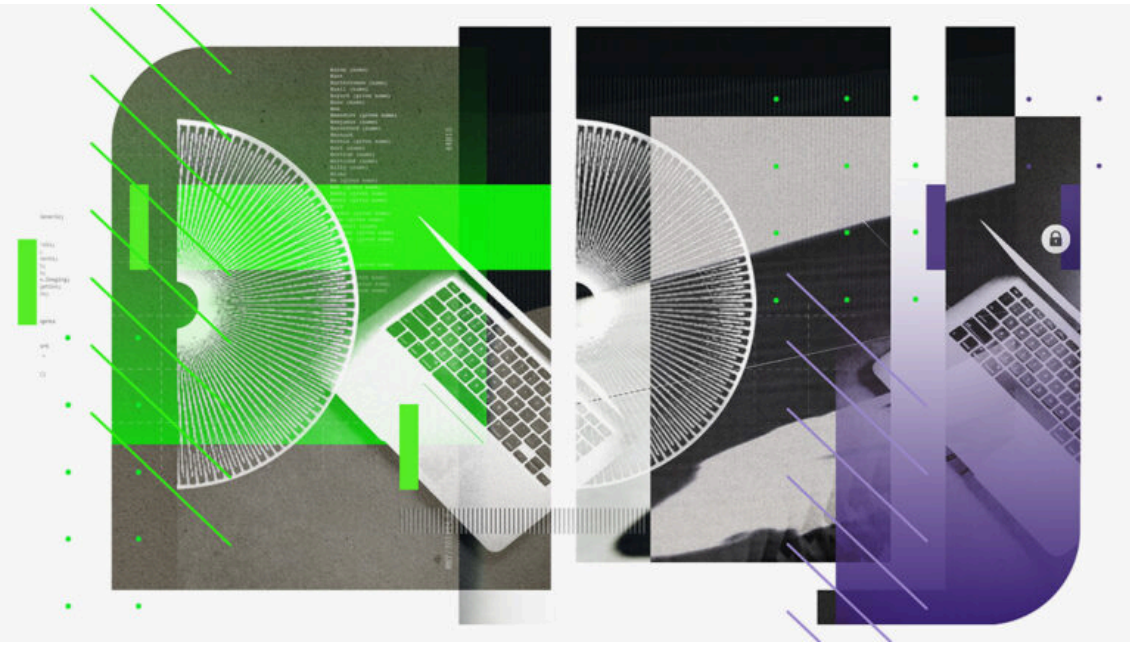
Harvard
Business
Review

Cybersecurity And Digital Privacy

The New Risks ChatGPT Poses to Cybersecurity

by Jim Chilton

April 21, 2023



Skizzomat

Summary. The FBI's 2021 Internet Crime Report found that phishing is the most common IT threat in America. From a hacker's perspective, ChatGPT is a game changer, affording hackers from all over the globe a near fluency in... **more**

When OpenAI launched their revolutionary AI language model ChatGPT in November, millions of users were floored by its capabilities. For many, however, curiosity quickly gave way to earnest concern around the tool's potential to advance bad actors' agendas. Specifically, ChatGPT opens up new avenues for hackers to potentially breach advanced cybersecurity software. For a

sector already reeling from a 38% global increase in data breaches in 2022, it's critical that leaders recognize the growing impact of AI and act accordingly.

Before we can formulate solutions, we must identify the key threats that arise from ChatGPT's widespread use. This article will examine these new risks, explore the needed training and tools for cybersecurity professionals to respond, and call for government oversight to ensure AI usage doesn't become detrimental to cybersecurity efforts.

AI-Generated Phishing Scams

While more primitive versions of language-based AI have been open sourced (or available to the general public) for years, ChatGPT is far and away the most advanced iteration to date. In particular, ChatGPT's ability to converse so seamlessly with users without spelling, grammatical, and verb tense mistakes makes it seem like there could very well be a real person on the other side of the chat window. From a hacker's perspective, ChatGPT is a game changer.

The FBI's 2021 Internet Crime Report found that phishing is the most common IT threat in America. However, most phishing scams are easily recognizable, as they're often littered with misspellings, poor grammar, and generally awkward phrasing, especially those originating from other countries where the bad actor's first language isn't English. ChatGPT will afford hackers from all over the globe a near fluency in English to bolster their phishing campaigns.

For cybersecurity leaders, an increase in sophisticated phishing attacks requires immediate attention, and actionable solutions. Leaders need to equip their IT teams with tools that can determine what's ChatGPT-generated vs. what's human-generated, geared specifically toward incoming "cold" emails. Fortunately, "ChatGPT Detector" technology already exists, and is likely to advance alongside ChatGPT itself. Ideally, IT infrastructure would integrate AI detection software,

automatically screening and flagging emails that are AI-generated. Additionally, it's important for all employees to be routinely trained and re-trained on the latest cybersecurity awareness and prevention skills, with specific attention paid to AI-supported phishing scams. However, the onus is on both the sector and wider public to continue advocating for advanced detection tools, rather than only fawning over AI's expanding capabilities.

Duping ChatGPT into Writing Malicious Code

ChatGPT is proficient at generating code and other computer programming tools, but the AI is programmed not to generate code that it deems to be malicious or intended for hacking purposes. If hacking code is requested, ChatGPT will inform the user that its purpose is to "assist with useful and ethical tasks while adhering to ethical guidelines and policies."

However, manipulation of ChatGPT is certainly possible and with enough creative poking and prodding, bad actors may be able to trick the AI into generating hacking code. In fact, hackers are already scheming to this end.

For example, Israeli security firm Check Point recently discovered a thread on a well-known underground hacking forum from a hacker who claimed to be testing the chatbot to recreate malware strains. If one such thread has already been discovered, it's safe to say there are many more out there across the worldwide and "dark" webs. Cybersecurity pros need the proper training (i.e., continuous upskilling) and resources to respond to ever-growing threats, AI-generated or otherwise.

There's also the opportunity to equip cybersecurity professionals with AI technology of their own to better spot and defend against AI-generated hacker code. While public discourse is first to lament the power ChatGPT provides to bad actors, it's important to remember that this same power is equally available to good actors. In addition to trying to prevent ChatGPT-related threats, cybersecurity training should also include instruction on how

ChatGPT can be an important tool in the cybersecurity professionals' arsenal. As this rapid technology evolution creates a new era of cybersecurity threats, we must examine these possibilities and create new training to keep up. Moreover, software developers should look to develop generative AI that's potentially even more powerful than ChatGPT and designed specifically for human-filled Security Operations Centers (SOCs).

Regulating AI Usage and Capabilities

While there's significant discussion around bad actors leveraging the AI to help hack external software, what's seldom discussed is the potential for ChatGPT itself to be hacked. From there, bad actors could disseminate misinformation from a source that's typically seen as, and designed to be, impartial.

ChatGPT has reportedly taken steps to identify and avoid answering politically charged questions. However, if the AI were to be hacked and manipulated to provide information that's seemingly objective but is actually well-cloaked biased information or a distorted perspective, then the AI could become a dangerous propaganda machine. The ability for a compromised ChatGPT to disseminate misinformation could become concerning and may necessitate a need for enhanced government oversight for advanced AI tools and companies like OpenAI.

The Biden administration has released a "Blueprint for an AI Bill of Rights," but the stakes are higher than ever with the launch of ChatGPT. To expand on this, we need oversight to ensure that OpenAI and other companies launching generative AI products are regularly reviewing their security features to reduce the risk of their being hacked. Additionally, new AI models should require a threshold of minimum-security measures before an AI is open sourced. For example, Bing launched their own generative AI in early March, and Meta's finalizing a powerful tool of their own, with more coming from other tech giants.

As people marvel at — and cybersecurity pros mull over — the potential of ChatGPT and the emerging generative AI market, checks and balances are essential to ensure the technology does not become unwieldy. Beyond cybersecurity leaders retraining and reequipping their workers, and the government taking a larger regulatory role, an overall shift in our mindset around and attitude toward AI is required.

We must reimagine what the foundational base for AI — especially open-sourced examples like ChatGPT — looks like. Before a tool becomes available to the public, developers need to ask themselves if its capabilities are ethical. Does the new tool have a foundational "programmatic core" that truly prohibits manipulation? How do we establish standards that require this, and how do we hold developers accountable for failing to uphold those standards? Organizations have instituted agnostic standards to ensure that exchanges across different technologies — from edtech to blockchains and even digital wallets — are safe and ethical. It is critical that we apply the same principles to generative AI.

ChatGPT chatter is at an all-time high and as the technology advances, it is imperative that technology leaders begin thinking about what it means for their team, their company, and society as a whole. If not, they won't only fall behind their competitors in adopting and deploying generative AI to improve business outcomes, they'll also fail to anticipate and defend against next-generation hackers who can already manipulate this technology for personal gain. With reputations and revenue on the line, the industry must come together to have the right protections in place and make the ChatGPT revolution something to welcome, not fear.

JC

Jim Chilton serves as CTO of Cengage Group, a global edtech company. He also currently serves as acting General Manager for the

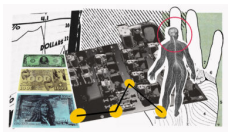
company's cybersecurity training business, Infosec.

Recommended For You

There's No Silver Bullet for Cybersecurity

Cybersecurity Needs to Be Part of Your Product's Design from the Start

Why Data Breaches Spiked in 2023



PODCAST
Why You (and Your Company) Need to Experiment with ChatGPT Now