THE WALL STREET JOURNAL.

How Hackers Can Up Their Game by Using ChatGPT

Cheryl Winokur Munk

June 5, 2023

Consumers, beware: AI chatbots like ChatGPT are likely to drive an increase in the use and effectiveness of online fraud tools such as phishing and spear-phishing messages.

In fact, it could already be happening. Phishing attacks around the world grew almost 50% in 2022 from a year earlier, according to Zscaler, a cloud-security provider. And, some experts say, artificial-intelligence software that makes phishing messages sound more believable are part of the problem. AI reduces or eliminates language barriers and grammatical mistakes, helping scammers impersonate a target's colleagues, friends or relatives.

"This new era is going to be worse than what we had before," says Meredith Broussard, research director at the New York University Alliance for Public Interest Technology. "And what we had before was really, really bad."

AI chatbots have exploded in popularity, with perhaps the best-known being ChatGPT, developed by the AI-research company OpenAI, a strategic partner of Microsoft. But dozens of chatbots, using what are referred to as large language models, are becoming more widely available and can closely mimic human communication based on data they amass. These models can be used for many purposes, such as helping office workers create routine memos more quickly. But they can also be used by criminals—to defraud victims, for instance, or to spread malicious viruses.

Telltale signs of a phishing attack have long included mistakes in grammar or spelling. But AI can give a phishing attack more credibility—and reach—not just because of its ability to generate fluent, grammatical messages in many languages, but also because of its ability to mimic the speaking or writing styles of individuals.

"The whole point with large language models is their ability to emulate what humans sound like," says Etay Maor, senior director of security strategy at Cato Networks, a cloud networking and security provider.

Thus, given the opportunity to learn the style in which a certain person writes emails and texts, Maor says, an AI program can be used to mimic communications from a company executive.

“It’s all about trust, and if I can make you think I’m one of you, you’re going to begin to do things with more trust and less skepticism,” says Roger Grimes, a computer-security professional with KnowBe4, a security-awareness training and simulated-phishing platform.

Using AI, Grimes says, criminals can quickly determine industry-specific terms that give them more ability to target companies such as hospitals, banks and fintech.

Targeted campaigns

AI’s usefulness in phishing and spear-phishing attacks doesn’t stop with its ability to mimic authentic human communication. The analytic skills of machine learning can also be useful in determining who best to target in an organization and how exactly to attack them.

Sean McNee, vice president of research and data at DomainTools, an internet intelligence company, offers a hypothetical example. Say an accountant at a company innocently posts on social media about his frustrations with a recent audit. AI could determine the accountant’s peers, his company’s reporting structure and who else at the company might be most susceptible to an attack. The attacker then could create a spear-phishing email purporting to be from the chief financial officer referring to a discrepancy in the audit and asking the recipient to open an attached spreadsheet that contains a virus.

Ramayya Krishnan, dean of Carnegie Mellon University’s Heinz College, recommends being proactive to protect against such attacks.

First, before acting on something, he says, people should always verify the legitimacy of the request through independent means. This means before clicking on a link or sending money, the recipient should call the individual through a familiar phone number or walk into the person’s office to confirm the request, Krishnan says.

Maintain a healthy dose of skepticism for everything you receive, Maor says. Ask yourself, why is my bank emailing me? Why is there a sense of urgency? Why is there an attachment to click on? It’s also advisable to hover over a link before clicking to see if it leads to an expected URL. “If you have some reason to think something is amiss, don’t click on it,” Maor says.

Other guardrails

Strong regulation of AI could also help, says Broussard, who is also an associate professor at the Arthur L. Carter Journalism Institute of New York University.

AI itself should also be enlisted to help identify malicious content with its origins in AI, says Dave Ahn, chief architect at Centripetal, a cybersecurity company. But first the models for doing so will have to evolve and the data will have to improve. Data on successful AI-based attacks will help cybersecurity experts train new models to identify malicious activity better, says Ahn.

Other possible security measures include giving users a way to distinguish their content as authentic. The use of hidden patterns known as “watermarks,” for instance, can be buried in AI-generated texts to help identify whether the words are written by a human or computer, Krishnan says. But the applicability of these tools is limited.

Says Krishnan, “We’re not near deploying them at scale where it’s a solution to the bad-actor potential we have today.”