ALM | LAW.COM

New York Law Journal

Conned by ChatGPT: The Growing Risks of AI-Powered Cyber Attacks

By David Owen and Alexa Moses
March 1, 2024

Artificial intelligence (AI) is transforming the things we can do and how we do them with amazing speed. Today almost any desired communication can be quickly and easily generated with the use of AI large language models (LLMs) like ChatGPT. As one would expect, the bots excel at technical writing. Chatbot LLMs can be prompted to write anything from error-free computer code to legal briefs (not to mention student homework). LLMs can even help fill in for more personal communications.

A recent episode of the animated television series South Park mocked the use of ChatGPT to generate plausible excuses and romantic text messages that were far more persuasive and compelling than anything the characters could come up with by themselves. Just as can be done in real life, the children repeatedly used AI-generated messages to fool and manipulate others, concerned only that their nefarious AI chat power remain a secret.

Unfortunately, the power of AI to do bad things is not a secret. Cyber criminals are already at work using these tools to groom and streamline so-called "social engineering" attacks to make them trickier and more plausible than ever before.

Adding to the challenge is the voluminous amount of private information that is being fed into these publicly available products. AI tools are designed to incorporate all the information they are fed and also

Photo: Production Perig via Adobe Stock

to return what they have received to anyone who asks in the proper manner. There is little apparent gatekeeping for what may be going in or coming out. Employees at companies around the world are entering private and sensitive data into ChatGPT with little concern for the possible consequences. As a result, the LLMs have themselves become sources of non-public information that can be used to sharpen attacks.

According to one study, nearly 5% of employees reported feeding confidential company data into ChatGPT. The actual number is likely much higher. In March of last year, OpenAI itself inadvertently leaked

its own customers' payment information by way of the chatbot, reporting "before we took ChatGPT offline on Monday, it was possible for some users to see another active user's first and last name, email address, payment address, the last four digits (only) of a credit card number and credit card expiration date." According to another recent report, leaked documents suggest that Amazon's new chatbot, Amazon Q, might be vulnerable to being tricked into revealing non-public customer information.

Security surrounding public LLMs is a significant challenge given their unpredictable behavior and unknown vulnerabilities. While ChatGPT and other LLMs have security features programmed into them to deter misuse and dissemination of private data, those measures can often be defeated in surprising ways. For instance, one security group persuaded ChatGPT to write convincing phishing emails by reassuring the bot that the messages were intended to be used for "employee awareness." Another disturbing report describes how ChatGPT can be used first to write phishing emails using perfect language and grammar, and then also to write functioning malware scripts to be delivered by way of the same email.

It is not difficult to see how all of this functionality can be combined into incredibly fast and effective social engineering attacks. The message will appear to come from a known merchant relating to a problem regarding your recent order. Everything about the communication will look authentic and familiar from the logo in the corner to the bland corporate language that one would expect—just another example of corporate automation and electronic efficiency.

Except it isn't, and the consequences can be devastating. Theoretically, any electronic communication is vulnerable to being attacked this way with just a bit of relevant information that often can be obtained from public or weakly-protected sources. Language and programming skills are no longer required.

With average losses from security events in the United States already exceeding $4 million per incident, it is fair to say that the number and reach of these attacks is only going to expand as it becomes easier for the criminals to fool people using these tools. Institutions will need to significantly enhance their defenses and security processes in response to the new and growing risks. Regular employee training and security exercises that include third-party vendors, contractors and others with access are vital to maintain a secure workplace. LLMs should be isolated from sensitive data and employees prohibited from sharing non-public data on a public LLM.

Multifactor authentication (MFA) must be implemented to cover everyone with any login access. MFA is an essential security feature that ensures that any credentials that are hacked will not grant any access by themselves. MFA protections should also be implemented in all payment processes.

Senior leadership will need to consistently emphasize the need to learn and follow all security processes, even when they are inconvenient or awkward.

Learning and following shifting security processes and passwords is already a time consuming challenge, but a consistent and regular training program remains the most effective way to ensure that people are prepared for the evolving dangers. "White hat" phishing exercises test employee awareness with imitation phishing emails to see how well employees respond to common efforts to trick them. Sharing performance metrics—and more pointed follow-up—can make the need for continuing vigilance less abstract.

Regular reminders also matter. One study that looked at the diminishing impact of phishing awareness training over time concluded that a four-month cycle may be a sweet spot for maintaining awareness.

Anti-phishing training is a baseline for all employees with access to sensitive systems or data—i.e., anyone with a login and password. According to the federal watchdogs at the Cybersecurity and Infrastructure Security Agency: "Employees should be able to identify the basic signs of phishing emails such as strange or unexpected requests, often using

alarming language or urging immediate action. These messages often appear to come from colleagues within the company or a trusted source. Malicious actors are improving their techniques all the time, so employees need to repeat training at regular intervals to learn about the latest scams."

The reference to regularly repeating and updating training appears to have the growing risks of AI LLMs particularly in mind. Future scams won't have the old red flags like a foreign prince or sketchy grammar to tip us off. The AI versions will be timely, slick and groomed by behavioral analytics.

In the broadest sense, the security threat posed by AI is the increasing need to distrust communications that may appear authentic or virtually authentic and which would have previously been thought to be generally reliable.

A security concept that has been evolving to address the problem head on is "zero trust"—i.e., treating all communications as if they are happening in a breach situation, always verifying authenticity by way of additional factors, and granting the least privileged access required. Originally coined by cyber analyst John Kindervag to describe a secure network architecture, the zero-trust concept has since expanded in scope and can be applied in principle to any process that needs to be secured against imposter fraud.

When an email box gets breached, an attacker gets a front-row seat and up to the minute information on everything passing through. As long as the attacker remains undetected, they can use that insight to learn how a company works and direct their focus to payment processes and the exfiltration of sensitive data. They may also gain the ability to impersonate the employee user in emails sent from the hacked box and fabricate correspondence ostensibly coming from people that the victim knows. Hackers have used that perch to trick victims into misdirecting astronomical sums. In theory, a zero-trust process should require sufficient checks and controls to prevent most if not all of these "man-in-the-middle" attacks.

Implementing zero-trust means adding inconveniences and potentially awkward double checks and significantly limiting institutional access. Barriers are put in place to ensure identities are fully trusted and verified. Passwords must be memorized and recycled more regularly. Employees will get additional passwords and authenticating devices and will face more elaborate hoops to get working again should they forget any of them. If implicit trust is the most expedient process available, zero-trust is the opposite.

As with any new security risk, the new risks posed by AI-powered social engineering attacks will inevitably require additional and inconvenient countermeasures to address them. On the bright side, AI can also help in the fight. Any attack process that has been automated is susceptible to being detected and stopped by an automated defense that knows what to look for. Armed with an understanding of attack patterns and features, defensive AI systems can quickly analyze huge volumes of data to detect suspicious activity and threats in real time.

But defensive systems can only work if they are put in place on a timely basis and constantly updated and patched for the most current threats and known vulnerabilities. Most often, people are the weakest link. Individuals and companies defending against these attacks have their work cut out for them.

David Owen *is a partner at Cahill Gordon & Reindel, where he advises clients on litigation, investigation and regulatory matters in various contexts, including securities law and privacy and cybersecurity law, among others.* **Alexa Moses** *is an associate in the firm's New York office, where she focuses her practice on intellectual property and data privacy law.*