Briefing: The New Artificial Intelligence (AI) Tools Driving Social Engineering Activity

CORVUS BY TRAVELERS

Portions of this article were previously published in the Travelers Cyber Threat Report

In the past few years there has been a persistent mismatch between the world-changing possibilities embodied by new AI tools and the reality of the as-yet-limited impact these tools are having on things like corporate financial results and employment numbers. It's been a similar story in the world of cybercrime: though it does not take a strong imagination to think of a scenario where threat actors use AI to increase their efficiency and capability at little cost, evidence of major changes in the patterns and habits of threat actors has been scarce.

That mismatch may be starting to break down. New research from the past quarter has shed light on the ways AI is making its way into the efforts of threat actors on several fronts. The following are a sample of the recent findings from security researchers around the industry, including at Travelers, on the frontier of AI usage.

AI-Driven Polymorphic Phishing

A notable evolution in phishing tactics is the rise of AI-powered polymorphic phishing – a technique that uses generative AI to craft and continuously modify phishing emails in real time. These polymorphic characteristics make it exceedingly difficult for conventional filters and static rule-based security systems to detect malicious content. In fact, Knowbe4, a security training firm, has predicted that by 2027, the traditional method of classifying phishing incidents into discrete campaigns may become obsolete, as polymorphic attacks will continue to evolve independently, rendering static categorizations ineffective.

GhostGPT: A New Frontier in AI-Enabled Phishing

One of the most notable tools fueling the recent spike in AI-enabled phishing is GhostGPT – an uncensored AI chatbot specifically tailored for cybercriminal use. Unlike mainstream AI platforms, GhostGPT operates no safeguards, allowing users to generate phishing emails, malware and exploit code on demand.

In January 2025, researchers from Abnormal Security identified GhostGPT as the source behind a rise in suspicious phishing emails targeting clients. In one test, the AI was prompted to generate a DocuSign phishing email, which it executed with a degree of realism and deception that made it suitable for immediate deployment in real-world attacks. This easy accessibility may lower the barrier to entry for less-skilled threat actors, accelerating the production and personalization of phishing lures across industries.

The Rise of Criminal AI Ecosystems

GhostGPT is part of a broader trend of malicious AI platforms that began emerging shortly after OpenAI released ChatGPT in late 2022. WormGPT, introduced in early 2023, was among the first AI models explicitly designed for malicious purposes. It was soon followed by other variants like WolfGPT and EscapeGPT.

These tools are not isolated developments. Researchers have now observed advertisements for GhostGPT on dark web forums, complete with subscription models, user guides and customer support – mirroring legitimate software-as-a-service (SaaS) offerings. This shift indicates that cybercriminals are building full-scale business infrastructures to support and monetize AI-driven attacks.

Even legitimate large language models (LLMs) are in the cross hairs. Research from Cisco reveals that fine-tuning LLMs for specific tasks significantly compromises their safety and security features. Models fine-tuned for domains like biomedicine, finance and law exhibited a much greater likelihood of generating harmful responses compared to their original versions. This vulnerability is exploited by cybercriminals by weakening guardrails and opening the door to jailbreaks, prompt injections and model inversion.

Deepfake-Driven Attacks

While it's still early in the evolution of deepfakes, they continue to be observed as tactics used for social engineering. For example, last quarter saw a phishing campaign that targeted YouTube creators using videos impersonating YouTube's CEO to steal credentials. There have also been reports of successful attacks on executives resulting in the fraudulent transfer of funds.

At Travelers, we took the opportunity to test this out for ourselves. Using widely available tools, we set out to see if we could develop a convincing deepfake of an executive. Using only a few minutes of audio from an interview given by a Travelers executive, coupled with a still image taken from the video of the interview, we were able to generate a reasonably convincing video-animated deepfake.

Gartner forecasts that by next year, 30% of enterprises will view identity verification and authentication solutions as unreliable when used alone, due to the prevalence of AI-generated deepfakes. It is anticipated that the use of deepfakes to target organizations will persist, manifesting through methods such as fraud, phishing, smishing or vishing.

Although identifying manipulated videos, images or voices can be challenging, recognizing unusual email addresses, links or phone numbers as potential red flags is a critical first step. Providing employees with cyber awareness training and keeping them informed about ongoing threats is essential for mitigating these risks.