Advanced attack vector

AI agents are opening the door to new hacking threats.

By THOMAS URBAIN

CYBERSECURITY experts are warning that artificial intelligence agents, widely considered the next frontier in the generative AI revolution, could wind up getting hijacked and doing the dirty work for hackers.

AI agents are programs that use artificial intelligence chatbots to do the work humans do online, like buy a plane ticket or add events to a calendar.

But the ability to order around AI agents with plain language makes it possible for even the technically non-proficient to do mischief.

"We're entering an era where cybersecurity is no longer about protecting users from bad actors with a highly technical skillset," AI startup Perplexity said in a blog post.

"For the first time in decades, we're seeing new and novel attack vectors that can come from anywhere."

These so-called injection attacks are not new in the hacker world, but previously required cleverly written and concealed computer code to cause damage.

But as AI tools evolved from just generating text, images or video to being "agents" that can independently scour the Internet, the potential for them to be commandeered by prompts slipped in by hackers has grown.

"People need to understand there are specific dangers using AI in the security sense," said software engineer Marti Jorda Roca at NeuralTrust, which specialises in large language model security.

Meta calls this query injection threat a "vulnerability". OpenAI chief information security officer Dane Stuckey has referred to it as "an unresolved security issue".

Both companies are pouring billions of dollars into AI, the use of which is ramping up rapidly along with its capabilities.

AI 'off track'

Query injection can in some cases take place in real time when a user prompt "book me a hotel reservation" – is gerrymandered by a hostile actor into something else – "wire US$100 (RM415) to this account."

But these nefarious prompts can also be hiding out on the Internet as AI agents built into browsers encounter online data of dubious quality or origin, and potentially booby-trapped with hidden commands from hackers.

Eli Smadja of cybersecurity firm Check Point sees query injection as the "number one security problem" for large language models that power AI agents and assistants that are fast emerging from the ChatGPT revolution.

Major rivals in the AI industry have installed defences and published recommendations to thwart such cyberattacks.

Microsoft has integrated a tool to detect malicious commands based on factors including where instructions for AI agents originate.

OpenAI alerts users when agents doing their bidding visit sensitive websites and blocks proceeding until the software is supervised in real time by the human user.

Some security professionals suggest requiring AI agents to get user approval before performing any important task – like exporting data or accessing bank accounts.

"One huge mistake that I see happening a lot is to give the same AI agent all the power to do everything," Smadja said.

In the eyes of cybersecurity researcher Johann Rehberger, known in the industry as "wunderwuzzi," the biggest challenge is that attacks are rapidly improving.

"They only get better," Rehberger said of hacker tactics.

Part of the challenge, according to the researcher, is striking a balance between security and ease of use since people want the convenience of AI doing things for them without constant checks and monitoring.

Rehberger argues that AI agents are not mature enough to be trusted yet with important missions or data.

"I don't think we are in a position where you can have an agentic AI go off for a long time and safely do a certain task," the researcher said.

"It just goes off track." – AFP