

Chatbots Are Becoming Really, Really Good Criminals

Matteo Wong

Earlier this fall, a team of security experts at the AI company Anthropic uncovered an elaborate cyber-espionage scheme. Hackers—strongly suspected by Anthropic to be working on behalf of the Chinese government—targeted government agencies and large corporations around the world. And it appears that they used Anthropic’s own AI product, Claude Code, to do most of the work.

Anthropic published its [report](#) on the incident earlier this month. Jacob Klein, Anthropic’s head of threat intelligence, explained to me that the hackers took advantage of Claude’s “agentic” abilities—which enable the program to take an extended series of actions rather than focusing on one basic task. They were able to equip the bot with a number of external tools, such as password crackers, allowing Claude to analyze potential security vulnerabilities, write malicious code, harvest passwords, and exfiltrate data.

Once Claude had its instructions, it was left to work on its own for hours; when its tasks were concluded, the human hackers then spent as little as a couple of minutes reviewing its work and triggering the next steps. The operation appeared professional and standardized, like any other business: The group was active only during the Chinese workday, Klein said, took a lunch break “like clockwork,” and appeared to go on vacation during a major Chinese holiday. Anthropic has said that although the firm ultimately shut down the operation, at least a handful of the attacks succeeded in stealing sensitive information. Klein said he could not provide further details, but that targets aligned with “strategic objectives of the Chinese government.” (A spokesperson for the Chinese embassy in Washington [told](#) *The Wall Street Journal* that its government “firmly opposes and cracks down on all forms of cyberattacks” and called such allegations by the United States “smear and slander.”)

[Read: The criminal enterprise behind that fake toll text](#)

We may now be in the “golden age for criminals with AI,” as Shawn Loveland, the chief operating officer at the cybersecurity firm Resecurity, put it to me. The recent hacking operation using Claude is just one of many examples: State-sponsored hacking groups and criminal syndicates are using generative-AI models for all manner of cyberattacks.

Anthropic, OpenAI, and other generative-AI companies proudly advertise AI’s ability to write code. But a boon for reputable businesses and software engineers is also one for cybercriminals. “Malware developers are developers,” Giovanni Vigna, the director of the NSF AI Institute for Agent-Based Cyber Threat Intelligence and Operation, told me—of course they’re going to take advantage of AI, just like everyone else. A student can use a chatbot to blast through their history homework, and a hacker can use it to speed through tasks that might otherwise take hours or days: writing phishing emails, debugging ransomware, identifying vulnerabilities in public codebases. Respected tech firms try to put safeguards in place to prevent their bots from being used to create malicious code, but they can be tricked; a user can pose as a participant in a cybersecurity competition, as experts at Google recently reported, which may lead the AI to comply with their requests.

OpenAI, Google, and Anthropic have uncovered Russian, Iranian, and Chinese hacker groups, among others, using their AI models to accelerate and scale their operations. A criminal enterprise or intelligence agency might typically have dozens or hundreds of skilled human hackers on their payroll, Vigna said. Now “suppose with the push of a button you can have a million of them—this is the power of AI.” AI models may not work at the level of a human developer, but their threat is already evident: A recent [experiment](#) by a team at UC Berkeley used AI agents to identify 35 new security holes in a group of public codebases. In other words, bots are able to find vulnerabilities that people miss.

Generative AI may be pushing us toward something like a worst-case scenario for basic cybersecurity. People are beginning to develop malware that can use large language models to write custom code for each hacking attempt, rather than using the same program for every machine or database targeted—a process that makes attacks much harder to detect, and one that security experts have been worried about “for 20-plus years,” Billy Leonard, an engineer in Google’s threat-analysis group, told me. Meanwhile, a digital black market for AI hacking tools is making even the most advanced techniques more and more accessible; less skilled hackers are able to launch much more effective attacks now than they would have been able to just a few years ago. The bots are making intrusions faster as well, perhaps so much so that by the time defense mechanisms kick in, “your attacker could be deep in your network,” Brian Singer, a cybersecurity expert at Carnegie Mellon University, told me.

And it’s not just that AI tools are powerful. In fact, another problem is that AI is actually ... kind of dumb. Businesses have rushed to deploy buzzy chatbots and AI agents, but these programs are themselves vulnerable to all sorts of clever and devastating attacks. “Nobody is really doing adequate threat modeling,” Loveland said—a company that rushes to put, say, customer-service bots in front of users may be opening up a new way for hackers to push malicious code and access users’ data or security credentials. On top of that, more and more software engineers (and hobbyists) are using AI to generate code, without taking the time (or even knowing how) to do basic security checks, which is introducing “a lot of new security vulnerabilities,” Dawn Song, a cybersecurity expert at Berkeley, told me.

[Read: Here’s how the AI crash happens](#)

IT professionals are also trying to leverage the technology for cybersecurity. Just as you might have 1 million virtual hackers, Vigna said, a company could create “millions of virtual security analysts” to look at your code—which he said could have disproportionate benefits to typically under-resourced IT experts. Instead of finding vulnerabilities to exploit, an AI model can find vulnerabilities to patch. Several cybersecurity experts told me the technology could be a boon for network defense in the long run. AI tools can offer the ability to audit large digital infrastructures, all the time, and at unprecedented speeds, Adam Meyers, the head of counter-adversary operations at the cybersecurity firm CrowdStrike, told me.

An all-out AI hacking arms race is afoot, and nobody can definitively say who will come out ahead. In the short term, the AI boom may well give cybercriminals the upper hand. Even before ChatGPT, attackers had an edge: Hackers have to discover only one vulnerability to succeed, while defenders have to miss only one to fail; hackers will rapidly try new methods, while businesses have to be slow and cautious. The better attackers get at using AI models, and the better the technology itself becomes, the harder intrusions will be to guard against. Then again, AI products that uncover new security flaws could also help patch those bugs. (And then those AI tools could be used by hackers to find security flaws in those patches. And so on.)

But no matter how fast an AI security tool can find a vulnerability, large companies and government agencies are far more risk-averse than hackers, Song said, because the smallest error could bring down an entire codebase or business—meaning, she said, that even if AI can quickly find bugs, defenders may remain slower to patch them. “Honestly, the last five to 10 years, cyberattacks have evolved, but the techniques to do these hacks have been somewhat consistent,” Singer said. “Now there’s kind of this paradigm shift,” and nobody can fully predict the fallout.