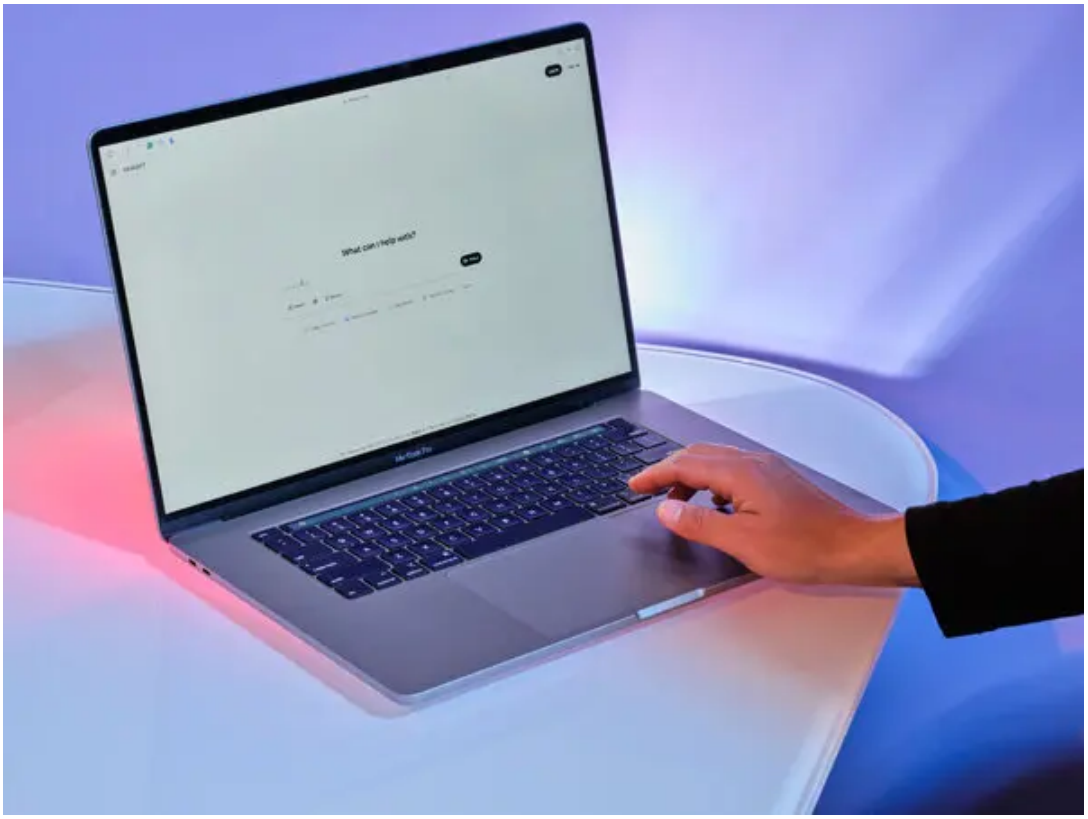
A.I. Is on Its Way to Upending Cybersecurity

Cade Metz, Kate Conger

With new systems from companies like Anthropic and OpenAI, hackers can attack with greater speed. The defense is more A.I.

Listen · 7:56 min



As tech companies prepare to release new and more powerful A.I. systems in the coming weeks, cybersecurity experts have become increasingly vocal in their warnings that A.I. technologies are fundamentally changing cybersecurity.Credit...Kelsey McClellan for The New York Times

Published April 6, 2026Updated April 7, 2026

Anthropic said late last year that state-sponsored Chinese hackers had used its artificial intelligence technology in an effort to infiltrate the computer systems of roughly 30 companies and government agencies around the world.

In a blog post, Anthropic said it was the first reported case of a cyberattack in which A.I. technologies had gathered sensitive information with limited help from human operators. Human hackers, the company said, handled about 10 to 20 percent of the work needed to conduct the attack.

Five months later, that remains the only known example of a cyberattack driven largely by an “A.I. agent” — technology that can write computer code and use software on its own. But as Anthropic and its chief rival, OpenAI, prepare to release new and more powerful A.I. systems, cybersecurity experts are increasingly vocal in their warnings that A.I. is fundamentally changing cybersecurity.

Technology from Anthropic, OpenAI, Google and other companies could allow hackers to identify security holes in computer systems far faster than in the past, vastly raising the stakes in the decades-long fight between hackers and the security experts guarding computer networks.

But like other tools from the long history of cybersecurity, the latest A.I. can be used for both offense and defense. As hackers deploy A.I. to break and steal, security experts are also leaning on A.I. to spot flaws in their systems — including some that had gone unnoticed for decades. The question is who finds the flaws first.

“This is the most change in the cyber environment, ever,” said Francis deSouza, the chief operating officer and president of security products at Google Cloud. “You have to fight A.I. with A.I.”

Since last year, the leading open source software projects — which provide the underlying infrastructure for sites and services across the internet — have been flooded with messages from people using A.I. to identify security holes.

Many of these so-called bug reports were erroneous, because of mistakes made by the A.I. systems. But in recent months, as A.I. has improved, they have started to identify legitimate bugs at a remarkable rate, and programmers have raced to make fixes.

“These A.I. models are augmenting what humans can do,” said Daniel Stenberg, who runs an important and popular open source project called Curl. “If you use these tools correctly, they can really raise your ability to find problems in software.”

In February, Anthropic said it had used its A.I. technologies to find over 500 so-called zero-day vulnerabilities — security holes that were unknown to software makers — in various pieces of commonly used open source software. The next month, a researcher at Anthropic revealed that he had used A.I. to find a serious security vulnerability in the core of the Linux operating system, which is software that powers much of the internet and is used in computer servers, cloud computing services, Android phones and Teslas.

The bug had existed, apparently undiscovered, since 2003.

Experts disagree on whether one side of this struggle has gained a significant advantage through A.I. And they are unsure how the battle will play out in the coming years. But most agree that the companies and governments that do not embrace the latest A.I. for defensive purposes will leave themselves enormously vulnerable.

Chatbots like Anthropic’s Claude and OpenAI’s GPT have become very good at writing computer code. These systems can help engineers create new software. They can use internet tools, like email programs and online calendars. And they can probe the weak points in software and online services, looking for security vulnerabilities.

Over the past several months, new A.I. tools like Anthropic’s Claude Code and OpenAI’s Codex — specifically made for coding — have helped developers create A.I. agents that can handle a wide variety of tasks largely on their own. That includes identifying and exploiting security holes in software.

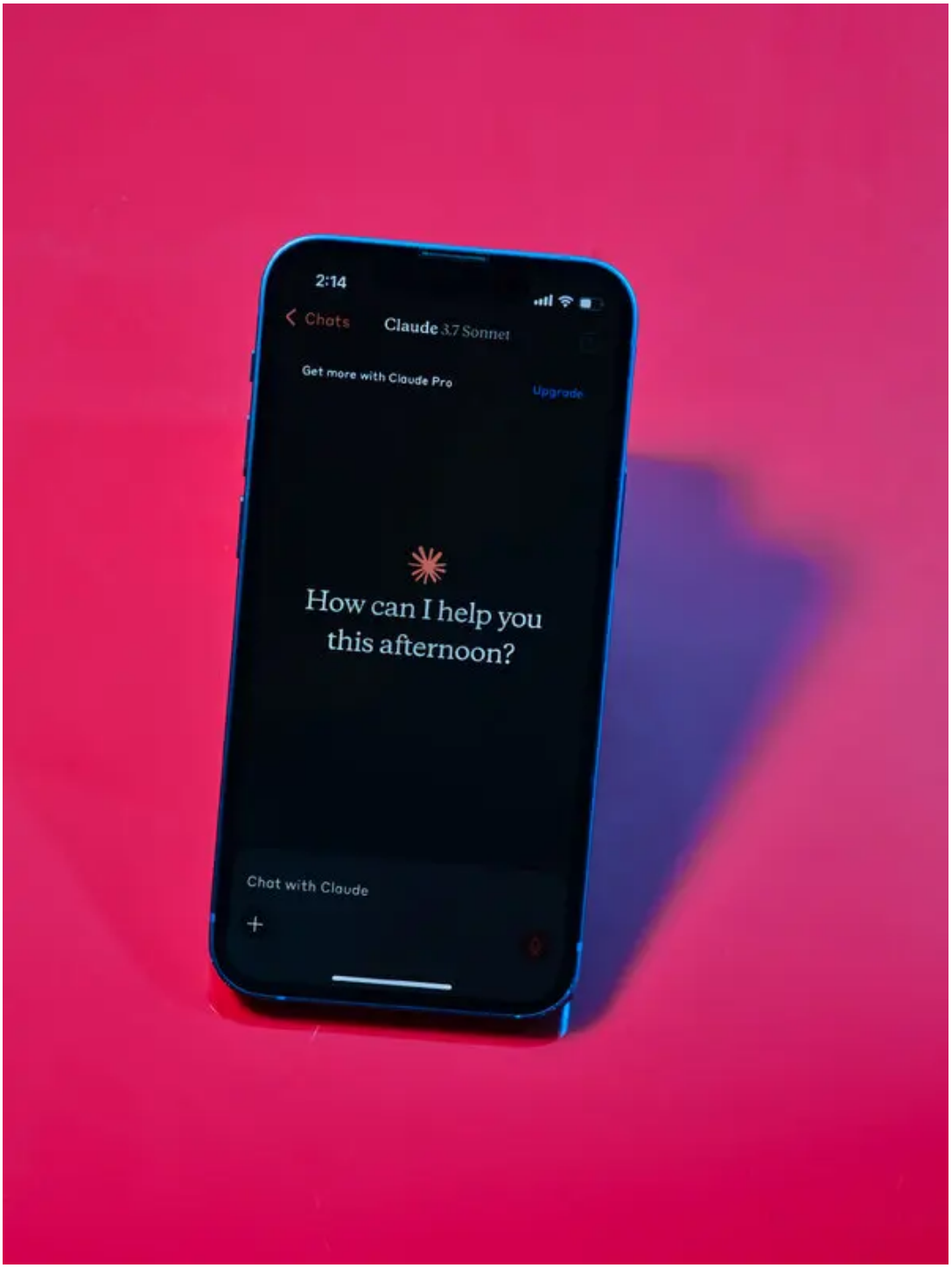
“Four or five months ago, we had a step change in what these systems could do,” said Zico Kolter, an OpenAI board member and a professor of computer science at Carnegie Mellon University who specializes in security and A.I.

A.I. is helping attackers in other ways. Some have used chatbots to draft phishing emails and ransom notes, cybersecurity experts said. Others have used A.I. to parse large quantities of stolen data and determine what information might be valuable. Without help from A.I., attackers could sometimes break into computer networks within minutes, Mr. deSouza said, but with the help of A.I., breaches can take just seconds.

Some hackers specialize in breaking into systems and then selling off their access to other attackers. Those handoffs used to take as much as eight hours, as hackers negotiated the sales and passed along the compromised entry points, Mr. deSouza said. Now that process has accelerated to about 20 seconds, he said, with hackers sometimes using A.I. agents to speed up the process.

Anthropic, OpenAI and other A.I. companies have tried to add guardrails to their tools to prevent them from being turned into cyberweapons. But attackers have been able to circumvent these barriers by telling the A.I. systems that they are not actually attacking.

Image



Anthropic has reported a cyberattack in which A.I. technologies gathered sensitive information with limited help from human operators.Credit...Kelsey McClellan for The New York Times

For instance, they will say are just playing “capture the flag” games — cybersecurity exercises that simulate real attacks and allow engineers to practice finding and exploiting vulnerabilities.

Some experts argue that the guardrails added by companies like Anthropic and OpenAI can actually provide an advantage to malicious attackers. Guardrails could cause a chatbot to deny help to a user trying to defend a system from an attack, they argue, but persistent hackers could be more diligent about finding vulnerabilities — and keeping those tricks to themselves.

“Claude is built with strong safeguards to prevent misuse of our models,” an Anthropic spokeswoman, Paruul Maheshwary, said in statement. “As the barriers to performing sophisticated cyberattacks continue to drop, we believe these protections are essential to preventing A.I. from being used as a tool for attackers.”

Although A.I. technologies have put new powers in the hands of offensive hackers, experts are divided over whether these tools give attackers an overall advantage over defenders.

Even after months of steady improvements, A.I. technologies are still flawed — which means they require the expertise of seasoned cybersecurity experts. In many cases, the tools are still limited by the skills of the people who use them.

“You still need a software architect in the loop with these systems,” Dr. Kolter said.

He and others argue that defenders have an advantage because they have the easier job. They just have to find the holes. Offensive hackers must both find and exploit the holes.

“It is easier to find a vulnerability than to meaningfully exploit it,” Dr. Kolter said.

(The New York Times sued OpenAI and Microsoft in 2023 for copyright infringement of news content related to A.I. systems. The two companies have denied those claims.)

Anthropic kicked off another round of discussion across the cybersecurity community last month when Fortune reported that the company had inadvertently published the contents of a blog post describing an A.I. system it has not yet released. The blog post said the technology represented another “step change” in A.I. performance.

Image

Nikesh Arora, the chief executive of Palo Alto Networks, warned in a blog post of coming threats.Credit...Jeenah Moon/Reuters

In the wake of the leaked blog post, Nikesh Arora, chief executive of the cybersecurity company Palo Alto Networks, published his own blog post warning companies and governments that they need to embrace the latest tools.

“The ability to identify vulnerabilities is going to be a lot better than we have ever seen before,” he said in an interview. “We have to get ready to fix these problems.”

Cade Metz is a Times reporter who writes about artificial intelligence, driverless cars, robotics, virtual reality and other emerging areas of technology.