Claude Mythos Preview Is Everyone's Problem

Matteo Wong

For the past several weeks, Anthropic says it secretly possessed a tool potentially capable of commandeering most computer servers in the world. This is a bot that, if unleashed, might be able to hack into banks, exfiltrate state secrets, and fry crucial infrastructure. Already, according to the company, this AI model has identified thousands of major cybersecurity vulnerabilities—including exploits in every single major operating system and browser. This level of cyberattack is typically available only to elite, state-sponsored hacking cells in a very small number of countries including China, Russia, and the United States. Now it's in the hands of a private company.

On Tuesday, the company officially announced the existence of the model, known as Claude Mythos Preview. For now, the bot will be available only to a consortium of many of the world's biggest tech companies—including Apple, Microsoft, Google, and Nvidia. These partners can use Mythos Preview to scan and secure bugs and exploits in their software. Other than that, Anthropic will not immediately release Mythos Preview to the public, having determined that doing so without more robust safeguards would be too dangerous.

For years, cybersecurity experts have been warning about the chaos that highly capable hacking bots could usher in. As a result of how capable AI models have become at coding, they have also become extremely good at finding vulnerabilities in all manner of software. Even before Mythos Preview, AI companies such as Anthropic, OpenAI, and Google all reported instances of their AI models being used in sophisticated cyberattacks by both criminal and state-backed groups. As Giovanni Vigna, who directs a federal research institute dedicated to AI-orchestrated cyberthreats, told me last fall: You can have a million hackers at your fingertips "with the push of a button."

Read: Chatbots are becoming really, really good criminals

Still, Mythos Preview appears to represent not an incremental change but the beginning of a paradigm shift. Until recently, the biggest advantage of AI-assisted hacking was not ingenuity, per se, so much as speed and scale. These bots could be as good as many human cybersecurity experts, but not necessarily better—rather, having an army of 1 million virtual, tireless hackers allows you to launch more attacks against more targets than ever before. Even Anthropic reports that its current state-of-the-art, public model, Claude Opus 4.6, was significantly less capable at autonomously finding cyber exploits. But Mythos Preview is different. According to Anthropic, the bot has been able to find thousands of software bugs that had gone undetected, sometimes for decades, a sophistication and speed of attack previously thought by many to be impossible. The model has found a nearly 30-year-old vulnerability in one of the world's most secure operating systems. The Anthropic researcher Sam Bowman posted on X that he was eating a sandwich in the park when he got an email from Mythos Preview: The bot had broken out of the company's internal sandbox and gained access to the internet.

The exact capabilities of Mythos Preview are hard to judge, because Anthropic has not released the model. Identifying a vulnerability is not the same as being able to exploit it undetected—in the same way that a robber can have the keys to a bank but still needs to deal with security cameras. And Anthropic surely stands to benefit from its opaque announcement: The company can claim to have developed an ultra-advanced model, while also appearing to act responsibly by preventing the worst-case cybersecurity scenarios. Indeed, the decision to not release Mythos Preview bolsters Anthropic's self-styled image as the AI industry's good guy. (Anthropic did not immediately respond to emailed questions about Mythos Preview.)

Of course, a move can be both strategic and conscientious. Should what Anthropic shared be remotely accurate, it heralds a troubling future. Anthropic has a tool that "could damage the operations of critical infrastructure and government services in every country on Earth," Dean Ball, a former AI adviser to the Trump administration, wrote this week. The ability to defend against such cyberattacks is integral to the basic functioning of society. And the ability to launch such attacks is integral to modern warfare. Anthropic may have just scaled its way into becoming a major geopolitical force.

Perhaps more concerning than the reported capabilities of Mythos Preview is that other companies are not far behind. OpenAI is reportedly set to release its own similarly powerful model to a select group of companies. It's very possible, even likely, that Google DeepMind, xAI, and AI firms in China are next. How scrupulous they will be is less clear. Even cheaper or open-source AI models from smaller companies could soon enable this sort of hacking—which would unsettle the basic security and privacy that undergird the modern internet.

Hacking bots are not the only domain through which a handful of AI companies are gaining tremendous influence. The technology has become crucial to military operations. Even as the Pentagon has engaged in a public feud with Anthropic, Claude was reportedly used in the bombing of Iran and, before that, the Venezuela raid in January. Last month, the Department of Defense signed a contract with OpenAI that very likely allows the government to use the firm's AI systems to enable unprecedented surveillance of U.S. citizens. (OpenAI has maintained that the Pentagon agreed not to use its products for domestic surveillance.) At the same time, bots from OpenAI, Anthropic, Google DeepMind, and beyond are becoming infrastructure: used by nearly all of the world's biggest businesses, schools, health-care systems, and public agencies. This is a large part of the reason that Iran has struck or threatened to strike Amazon and OpenAI data centers in the Middle East—the facilities are high-impact targets on par with the oil fields that Iran has also targeted. Meanwhile, so much money is pouring into the AI boom that these companies are functionally holding the global economy hostage.

In other words, AI companies are remaking the world. Consider how Elon Musk's network of Starlink satellites has allowed him to repeatedly tip the scales in Russia's invasion of Ukraine. Generative AI offers even more possibilities. These companies can or could soon have the capability to launch major cyberattacks, conduct mass surveillance, influence military operations, cause huge swings in financial and labor markets, and reorient global supply chains. In theory, nothing governs these companies other than their own morals and their investors. They are developing the power to upend nations and economies. These are the AI superpowers.