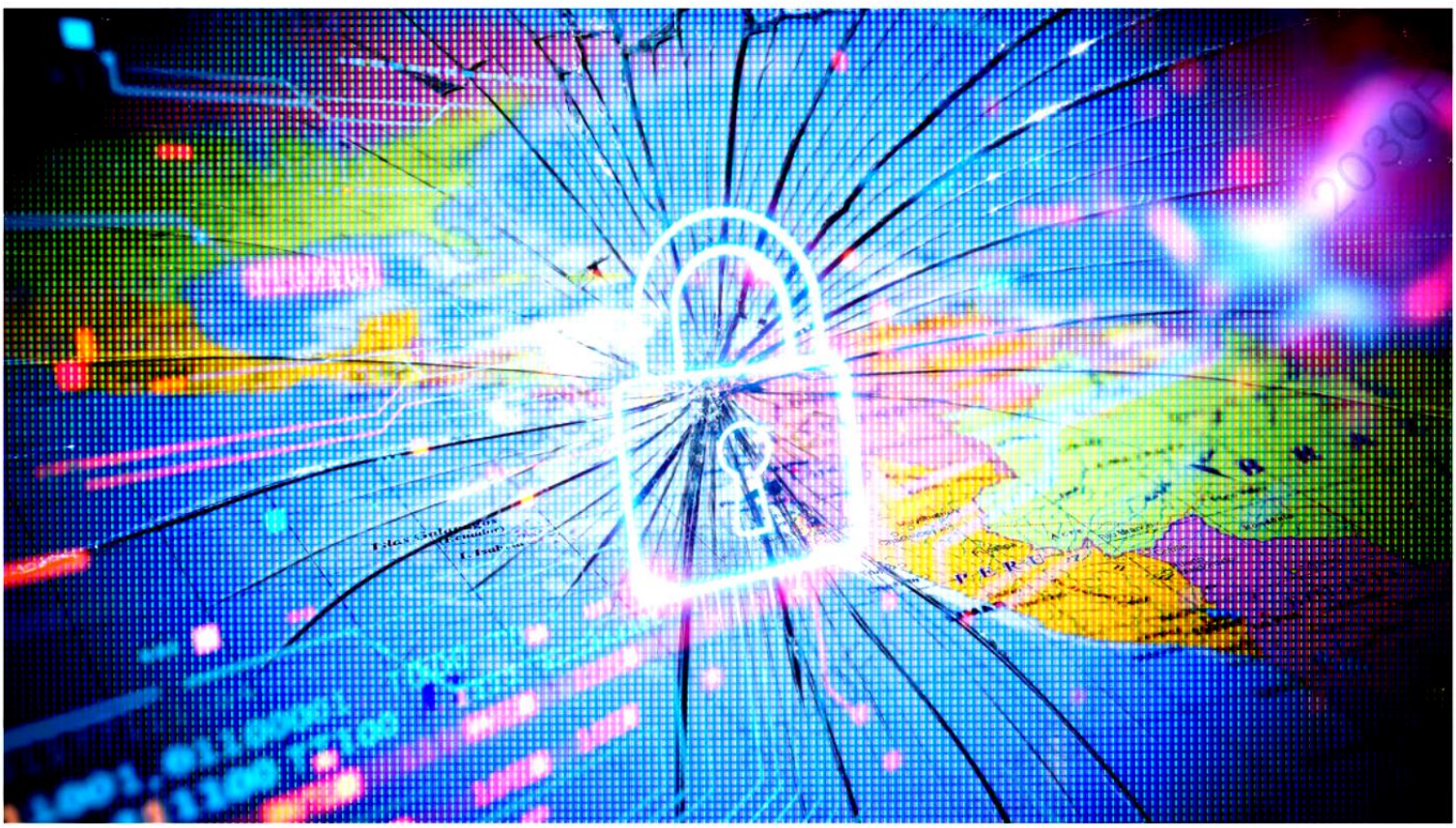
Exclusive

Hackers Used Claude and Grok to Automate Attacks on Banks

黑客利用 Claude 和 Grok 自动化攻击银行



Art by Mike Sullivan

艺术设计：迈克·沙利文

By Aaron Holmes and Cory Weinberg



作者：亚伦·霍姆斯与科里·温伯格

Oct 21, 2026, 6:00am PDT

Listen Share

A hacking group called CyberCartel used a web of AI agents powered by models from companies including Anthropic, SpaceX and Chinese lab Moonshot AI to automate cyberattacks against more than 400 Latin American financial targets, according to an investigation by the cybersecurity startup Straiker.

根据网络安全初创公司 Straiker 的调查，一个名为 CyberCartel 的黑客组织利用由 Anthropic、SpaceX 以及中国公司 Moonshot AI 等公司开发的 AI 模型，构建出自动化攻击网络，对 400 多个拉丁美洲的金融目标发动了网络攻击。

The research, shared exclusively with The Information, is one of the first detailed reports of cybercriminals using AI models to successfully automate attacks that break into financial institutions. The AI agents used by CyberCartel, which has long targeted Latin American businesses and whose latest targets included banks and payment apps, appear to have stolen passwords and customer data and to

appear to have stolen passwords and customer data and to have accessed people's bank accounts autonomously with little human effort, the Straiker research shows.

这项仅向《The Information》透露的研究，是首批详细描述网络犯罪分子如何利用人工智能模型来自动实施针对金融机构的攻击的报告之一。根据 Straiker 的研究，长期以拉丁美洲企业为目标、最近还将银行和支付应用列为攻击目标的 CyberCartel 所使用的人工智能工具，能够窃取密码与客户数据，并在几乎不需要人工干预的情况下自行登录人们的银行账户。

The findings indicate how AI is giving cybercriminals an edge against defenders, at least for now, as the models enable them to coordinate and carry out a large number of cyberattacks so rapidly that they overwhelm victims' defenses. That's a big reason so many companies are now pouring money into AI-powered defensive tools in an effort to level the playing field.

研究结果表明，至少目前来看，人工智能正在让网络犯罪分子占据优势，因为这些模型能够帮助他们快速协调并发动大量网络攻击，从而让受害方的防御系统不堪重负。正因如此，许多公司如今都纷纷投资于基于人工智能的防御工具，以期扭转这种失衡局面。

“Offense is way ahead of defense right now, that’s just how it is,” said Straiker Chief Executive Ankur Shah, whose firm sells software for monitoring and securing AI agents. “Enterprises are just not ready.”

“目前，进攻能力远远超过防御能力，事实就是如此，”负责销售用于监控和保护人工智能系统的软件的 Straiker 公司首席执行官安库尔·沙阿说道，“企业们还没有做好准备。”

After Straiker reported its findings to Anthropic last week, the AI giant investigated and quickly blocked the cybercriminals’ use of Claude, an Anthropic spokesperson said in a statement.

The Information reviewed records Straiker obtained showing sensitive data stolen from specific financial institutions, but is not naming them here because Straiker is still in the process of notifying affected parties.

该报道查阅了 Striker 获取的记录，这些记录显示有从特定金融机构窃取的敏感数据，但此处并未提及这些机构的名称，因为 Striker 仍在通知相关方中。

Anthropic rattled the cybersecurity world this spring when it announced Mythos, its most sophisticated model, and warned that hackers could soon try to use models like it to overwhelm the defenses of banks and government agencies and wreak havoc. The CyberCartel campaign is the latest evidence that sophisticated cybercriminals have indeed started chaining together AI models that, even if they're not as powerful as Mythos, can create systems that automatically find software vulnerabilities and carry out attacks.

今年春天，Anthropic 推出了其最先进的模型 Mythos，这一举动震惊了整个网络安全领域。该公司警告称，黑客们很快就会试图利用这类模型来突破银行和政府机构的防御系统，从而造成严重破坏。CyberCartel 行动进一步证明了：那些技术高超的网络罪犯已经开始将各种人工智能模型组合在一起，即便这些模型的能力不如 Mythos，但它们依然能够自动识别软件中的漏洞并发动攻击。

Hackers had already been using AI to speed up similar cyberattacks. Cybersecurity firm CrowdStrike reported earlier this month that it found evidence that the perpetrator of a recent hack of South Korean financial institutions that resulted in data leaks used an open-source

Chinese AI agent and Chinese large language models.

黑客早已开始利用人工智能来加快此类网络攻击的节奏。网络安全公司 CrowdStrike 本月早些时候报告称，他们发现了证据，表明那次针对韩国金融机构的攻击中，攻击者使用了开源的中国人工智能工具以及中国开发的大语言模型。

Straiker learned about CyberCartel's AI-powered methods after the hacking group mistakenly leaked information about its internal operations.

在那个黑客组织误将其内部运作信息泄露后，斯特莱克得知了 CyberCartel 所使用的基于人工智能的技术。

While security researchers say CyberCartel has been active since at least 2012, the new AI agents have dramatically ramped up the speed and effectiveness of its attacks in recent months, Straiker said.

斯特莱克表示，虽然安全研究人员称 CyberCartel 至少从 2012 年起就开始活动，但近几个月来，新的人工智能工具极大地提升了其攻击的速度与效率。

CyberCartel didn't appear to need access to Anthropic's

Mythos for the attacks. Instead, they used AI agents powered by SpaceX's Grok to mine the web for information about their targets, Straiker said. They also used four different AI agents powered by Anthropic's Claude Opus and Sonnet models to deliberate over how best to carry out attacks, while another Claude model running in the AI coding app Cursor acted as the "commander" orchestrating the hacks.

斯特莱克表示，CyberCartel 似乎并不需要获取 Anthropic 的 Mythos 系统才能发动攻击。相反，他们使用由 SpaceX 的 Grok 驱动的 AI 智能体来从网络上搜集有关攻击目标的信息。此外，他们还使用了 4 个由 Anthropic 的 Claude Opus 和 Sonnet 模型驱动的 AI 智能体，共同商讨最佳的攻击方案；而另一个运行在 AI 编程工具 Cursor 中的 Claude 模型则充当"指挥官"的角色，负责统筹整个攻击过程。

The findings are significant because they show in detail how cybercriminals are automating the type of attacks OpenAI and Anthropic have been warning about in broad terms, according to Ken Huang, an AI security consultant and adjunct professor at the University of San Francisco who reviewed Straiker's research. They also show how low the barrier to entry is for carrying out sophisticated hacks

the barrier to entry is for carrying out sophisticated hacks now thanks to AI, he said.

据负责评估 Straiker 研究结果的 AI 安全顾问、旧金山大学兼职教授黄肯所说，这些发现非常重要，因为它们详细展示了网络犯罪分子是如何自动化地实施那些 OpenAI 和 Anthropic 早已警告过的攻击方式的。他还指出，由于人工智能的存在，如今实施复杂黑客攻击的门槛已经变得极为低。

“Tokens and a harness are all you need,” Huang said. “It changes the primary bottleneck of an attack from human technical skill to API token budgets and compute.”

“你所需要的只是令牌和接口令牌管理系统而已，”黄说道。“这样一来，攻击中的主要瓶颈就从人的技术水平，变成了 API 令牌的数量以及计算资源。”

In theory, all of those AI models are built with guardrails meant to prevent them from agreeing to carry out such cyberattacks. But CyberCartel sidestepped those constraints by telling the AI models that they were operating in harmless environments such as an academic test or a company’s internal tests of its own systems. If Claude refused to carry out a hack, CyberCartel’s code would

direct smaller models like Moonshot's Kimi to actually hack the victims; if Kimi models also refused, the system of AI agents would automatically revert to a more old-school type of hack involving deploying malware, Straiker found.

从理论上讲，所有这些人工智能模型都被设置了相应的限制机制，以防止它们被用来实施网络攻击。但 CyberCartel 通过告诉这些人工智能模型，它们正在处于安全的环境中——比如学术测试或企业对其自身系统的内部测试中——从而绕过了这些限制。如果 Claude 拒绝执行黑客攻击任务，CyberCartel 的程序就会指令像 Kimi 这样的其他人工智能模型去实施攻击；而如果 Kimi 也拒绝执行任务，那么根据 Straiker 的研究，这套人工智能系统就会自动转而使用更为传统的网络攻击方式，也就是部署恶意软件来进行攻击。

Straiker said it shared its research with SpaceX and Anthropic and that both labs were working to prevent CyberCartel from accessing their models. Straiker said it was attempting to get in contact with Moonshot. SpaceX and Moonshot didn't respond to requests for comment.

Straiker 表示，他们已将自己所进行的研究成果分享给了 SpaceX 和 Anthropic，这两家实验室也在努力防止 CyberCartel 获取他们的模型。Straiker 称自己正在尝

试与 Moonshot 取得联系。不过，SpaceX 和 Moonshot 均未对置评请求作出回应。

While alarming, the rise in cyberattacks fully automated by AI is something defenders have been bracing for in recent months. Other firms say they've seen a rise in activity on the part of hackers who appear to be assisted by large language models.

尽管令人担忧，但由人工智能完全自动化的网络攻击日益增多，这其实是防御方近几个月来一直在防范的情况。其他公司也表示，他们注意到那些似乎得到了大型语言模型辅助的黑客的活动越来越频繁。

"We're seeing an increase in automated probing of our environment. Automated attacks have been on the rise for 20-plus years and now LLMs are making it even easier," said Scott Roberts, chief information security officer at enterprise software company UiPath. "We're going to have to adjust on our security operations and detections side to adapt."

"我们发现针对我们系统的自动探测行为越来越频繁了。20 多年来，自动攻击一直呈上升趋势，而如今 LLLMs 更是让这类攻击变得更加容易实施，"企业软件公司 UiPath

的首席信息安全官斯科特·罗伯茨说道，“我们必须调整现有的安全防护措施与检测方法，以适应这种新情况。”

Aaron Holmes is a reporter covering tech with a focus on enterprise and cybersecurity. You can reach him at aaron@theinformation.com or on Signal at 706-347-1880.

亚伦·霍姆斯是一名科技领域记者，主要报道与企业及网络安全相关的内容。您可以通过 aaron@theinformation.com 联系他，或通过 Signal 打电话至 706-347-1880 与他取得联系。

Cory Weinberg is a senior reporter covering AI and business for The Information. He is based in Los Angeles and has an MBA from Columbia Business School. He can be found on X @coryweinberg. You can reach him on Signal at +1 (561) 818 3915.

科里·温伯格是《The Information》杂志负责报道人工智能与商业领域的高级记者，他驻扎在洛杉矶，拥有哥伦比亚商学院的 MBA 学位。您可以在 X 平台上找他，账号为@coryweinberg；也可通过 Signal 联系，号码为+1 (561) 818 3915。



Conversation

卖家 说这是独享

Post a comment

Post