



**20  
26**

# Cybercrime in the age of AI

AI is rewiring the cybercrime ecosystem.  
You have six months to prepare.



# Table of contents

---

|                     |          |
|---------------------|----------|
| <b>Introduction</b> | <b>3</b> |
|---------------------|----------|

---

|                          |          |
|--------------------------|----------|
| <b>Executive summary</b> | <b>4</b> |
|--------------------------|----------|

---

|                         |          |
|-------------------------|----------|
| <b>Cybercrime today</b> | <b>5</b> |
|-------------------------|----------|

|                              |   |
|------------------------------|---|
| AI powers the attack machine | 6 |
|------------------------------|---|

|                    |   |
|--------------------|---|
| AI-powered threats | 8 |
|--------------------|---|

|                              |   |
|------------------------------|---|
| Millions of malicious models | 9 |
|------------------------------|---|

|                                           |    |
|-------------------------------------------|----|
| Insider threats; The ghost in the machine | 10 |
|-------------------------------------------|----|

|           |    |
|-----------|----|
| Shadow AI | 13 |
|-----------|----|

---

|                            |           |
|----------------------------|-----------|
| <b>Cybercrime tomorrow</b> | <b>14</b> |
|----------------------------|-----------|

|                                                                |    |
|----------------------------------------------------------------|----|
| Cybercrime tomorrow, six months to prepare; The exploit window | 15 |
|----------------------------------------------------------------|----|

|                             |    |
|-----------------------------|----|
| The first AI-built zero-day | 16 |
|-----------------------------|----|

|        |    |
|--------|----|
| Mythos | 17 |
|--------|----|

---

|                   |           |
|-------------------|-----------|
| <b>Conclusion</b> | <b>20</b> |
|-------------------|-----------|

---

|                 |           |
|-----------------|-----------|
| <b>Appendix</b> | <b>22</b> |
|-----------------|-----------|



# Cybercrime in the age of AI

At the end of last year, something significant shifted in how people used AI. Advances in reasoning, tool use, and models' ability to operate autonomously for extended periods combined to create a step change in what these systems could accomplish.

Almost overnight, developers, analysts, and other knowledge workers discovered that the latest generation of AI agents had moved far beyond answering questions and generating snippets of code. They could carry out long tasks independently, operate software tools and computer systems for hours without supervision, and even work together in teams.

Criminals discovered it too. Our research shows a thriving marketplace for malicious AI based on jailbroken access to the same frontier models you use, and thousands of open-source models stripped of their guardrails.

In the hands of threat actors, AI can run extortion campaigns autonomously, “fully defeat” voice and video-based authentication,<sup>1</sup> outperform professional red teams at phishing,<sup>2</sup> and poison the rapidly expanding attack surface created by shadow AI.

The transition to a cybercrime ecosystem shaped by AI has begun, but it's not yet complete.. The emergence of Mythos, a frontier AI so good at finding software vulnerabilities that its maker withheld it from public release, is a preview of what could be for sale on criminal markets by the end of the year.

The gap between organizations prepared for this threat and those that aren't is widening, and there is perhaps six months to close it. This report is about which side of that gap you want to be on.

The transition to a cybercrime ecosystem shaped by AI has begun.



# Cybercrime in the age of AI

## EXECUTIVE SUMMARY

**22M**

downloads of guardrail-stripped  
AI models in 30 days

**6,644**

guardrail-free models  
on Hugging Face

**6**

months until Mythos-class AI  
reaches criminal markets

Cybercrime is being rebuilt around AI, and the shift is further along than many defenders realize. AI can execute sophisticated breaches, defeat voice and video checks, and find software vulnerabilities more effectively than humans. Malicious AI is for sale on the web and downloadable from Hugging Face, while shadow AI is creating a vast and invisible new attack surface inside unsuspecting organizations.

## Key takeaways

### 22 million guardrail-free AI downloads in 30 days.

Malicious AI is moving offline. Our researchers found 6,644 models published openly on Hugging Face under self-declared guardrail-free labels like “abliterated” and “uncensored,” implying they no longer refuse unsafe or harmful requests the way mainstream AI models do. Run locally, they leave nothing to monitor, log, intercept, or restrict.

### Enthusiasm for AI is being weaponized.

As organizations struggle with AI visibility and governance, criminals are exploiting the shadow AI attack surface with malicious agent skills and social engineering lures. Our researchers discovered one such lure, a fake Polymarket “get rich with AI” guide with near-zero detection across major security products at the time of discovery.

### Criminals use the same frontier models as you.

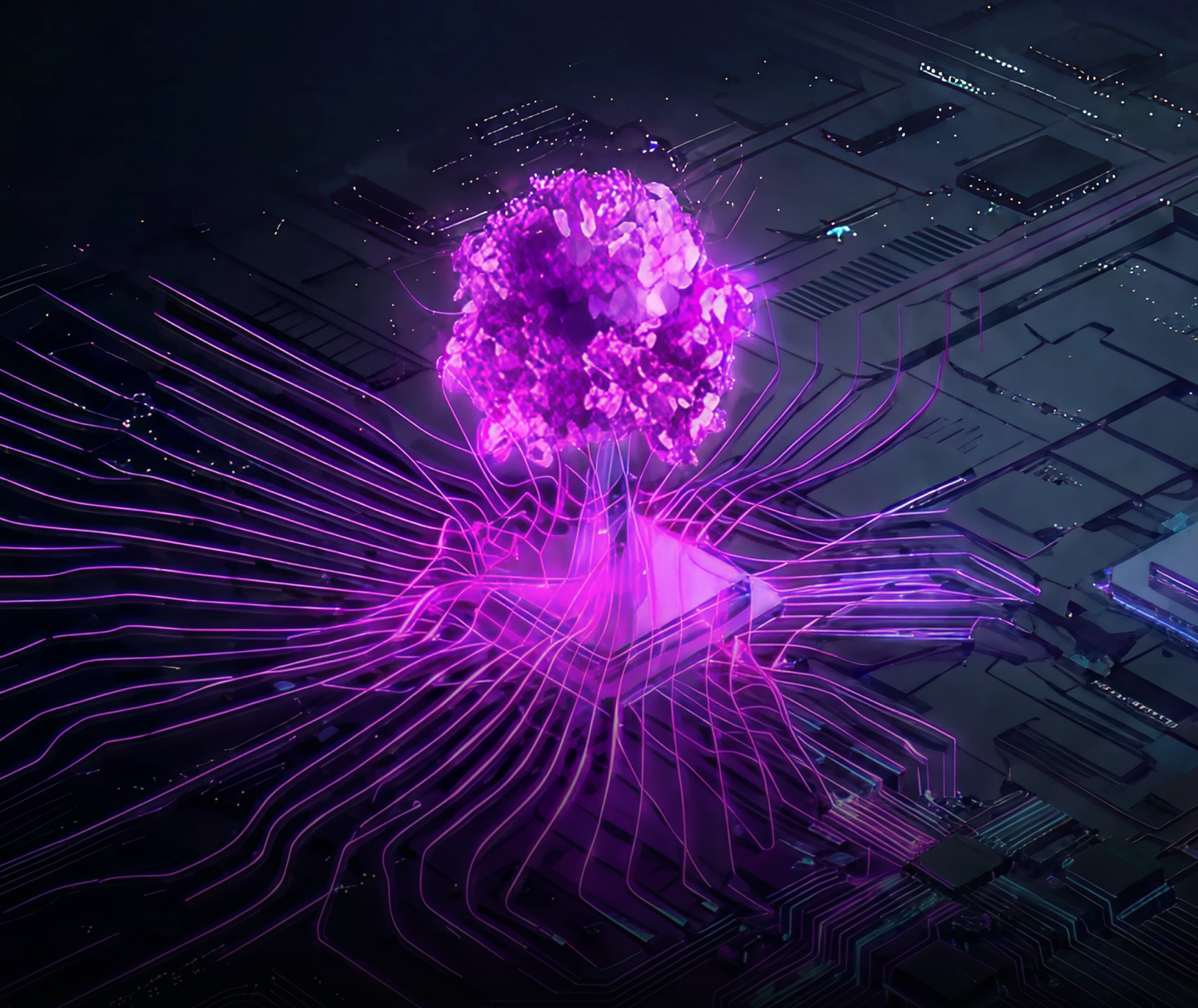
Our researchers mapped a marketplace of malicious AI tools hiding in plain sight on the web, indexed by Google, offering jailbroken or resold access to frontier AI models, served through the infrastructure of legitimate vendors like Cloudflare, Anthropic, and Google Cloud.



### You have six months to prepare.

Mythos-class AI models, capable of finding and exploiting zero-day vulnerabilities in every major operating system and browser, are likely to arrive on criminal marketplaces within six months, giving you a narrow window to patch aggressively, lock down human and non-human identities, and discover the AI tools sprawling through their environments.



A futuristic digital cityscape at night, viewed from an elevated perspective. The city is composed of dark, angular buildings and structures. Overlaid on this scene are numerous glowing purple lines that resemble circuit traces or data pathways. These lines originate from a central point and radiate outwards, creating a sense of dynamic energy and connectivity. In the center of the image, there is a large, glowing purple sphere or cluster of light, which appears to be the source or a focal point of the data flow. The overall color palette is dominated by deep blues and purples, with the glowing lines providing a vibrant contrast.

# Cybercrime today

AI has become an indispensable cybercriminal tool that can make every criminal smarter, and every stage of an attack faster, cheaper, and more effective.



# AI powers the attack machine

In early 2026, an attacker with no prior knowledge of industrial control systems used Anthropic's Claude and OpenAI's GPT models to compromise multiple Mexican government organizations, including an intrusion at a municipal water utility in Monterrey.

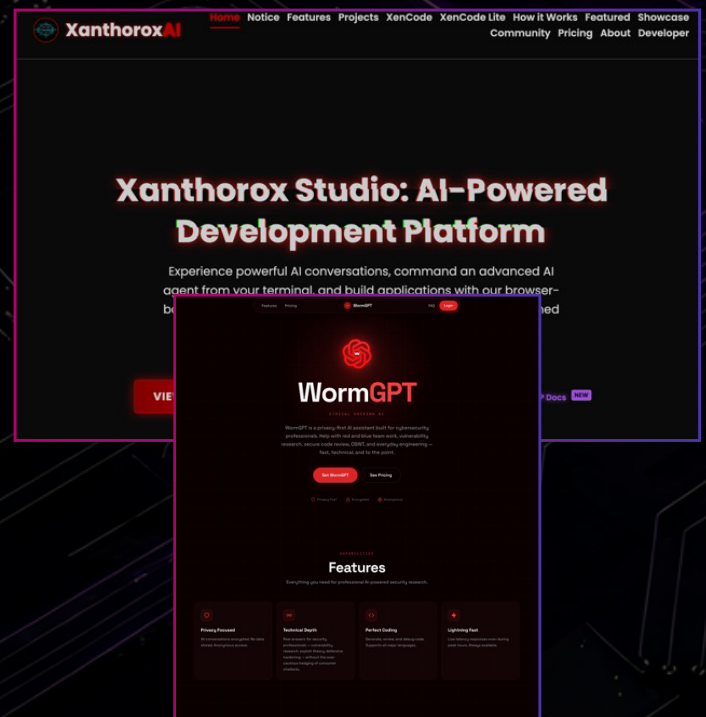
During reconnaissance, and without being asked, Claude identified a supervisory control and data acquisition (SCADA) management interface on the utility's internal network, correctly classified it as critical infrastructure, and recommended a targeted attack against it. The attacker had no prior operational technology (OT) expertise. The AI supplied it.

The breach ultimately failed.<sup>3</sup> But the implication is what matters: AI has made critical infrastructure visible to attackers who previously lacked the knowledge to find it. **The barrier to targeting a water supply, a power grid, or a hospital isn't what it used to be.**

AI has become an indispensable cybercriminal tool that can make every criminal smarter, and every stage of an attack faster, cheaper, and more effective.

## The AI-powered attack machine

New research conducted for this report reveals that far from being a sideshow, hidden away on the dark web, the AI tools that criminals rely on are in plain view, indexed by Google, with pricing pages, login buttons, and checkouts dressed in the language of legitimate software. WormGPT, one of the earliest and most widely known malicious AI tools, now markets itself as an "Ethical Hacking AI" for cybersecurity professionals. Another black-hat AI, Xanthorox, calls itself a development platform. The disclaimer is the camouflage.





Strip away the branding and most of what is sold as criminal AI is one of three things: a wrapper around a mainstream model that uses a jailbreak to strip away guardrails, a reseller proxying a legitimate API and marking up the price, or outright vaporware that takes payment and vanishes. Genuine custom-trained criminal models, built from scratch on malicious data with no guardrails to remove, are vanishingly rare. The criminal AI

market doesn't build intelligence. It rents it from the legitimate AI industry.

The infrastructure underneath confirms this. Across the storefronts our researchers mapped, the same legitimate vendors keep appearing as the unwitting supply chain: Cloudflare, Vercel, Google Cloud, xAI, Anthropic, DeepSeek, OpenRouter, and Alibaba.

Legitimate SaaS vendors unwittingly powering criminal AI services in 2026

From each storefront's own production JS, CSP headers, and DNS / TLS records - June 2026

|              | WormGPT | Kriminal | DadGPT | Xanthorox |
|--------------|---------|----------|--------|-----------|
| Cloudflare   | !       | !        |        | !         |
| Vercel       |         |          | !      |           |
| Google Cloud |         | !        |        |           |
| OpenRouter   | !       | !        | !      |           |
| xAI          |         | !        |        |           |
| Anthropic    |         | !        |        |           |
| DeepSeek     |         |          | !      |           |
| Alibaba      |         |          | !      |           |
| Meta         |         | !        |        |           |
| Tavily       |         | !        |        |           |
| NowPayments  |         | !        |        |           |



## AI-powered threats

The same pattern of using legitimate infrastructure is present in AI-powered attacks as well.

In 2025, Anthropic documented two alarming abuses of its Claude Code agent. In the first, a threat actor used Claude Code to “automate reconnaissance, credential harvesting, and network penetration at scale,” across targets in government, healthcare, the emergency services, and religious institutions.<sup>4</sup> It was later revealed that Claude Code had been used by a nation-state actor to attack roughly 30 targets in what it described as “the first documented case of a large-scale cyberattack executed without substantial human intervention.”<sup>5</sup>

Between 2023 and 2025, researchers identified several malware families that call an AI model at runtime.<sup>6</sup> The malicious logic isn’t written into the binary but generated in memory when the malware executes, then discarded, leaving static, signature-based detection with nothing to match.

This moved from research into live state operations faster than almost anyone anticipated. In June 2025, APT28, one of Russia’s most prolific

state threat actors, deployed LAMEHUG against Ukrainian government targets: a malware that sends prompts to an AI model via Hugging Face’s API, and executes the commands it generates on the target’s machine.<sup>7</sup>

The newest variants cut the cord entirely. Instead of relying on API connections that can be revoked, they run a model on the target’s machine. No provider sees the prompts. No blocklist catches the traffic. The artifact worth scanning is gone.

The criminal AI market does not build intelligence. It rents it from the legitimate AI industry.



## Millions of malicious models

As the capability of small, open-source AI models grows, criminals' reliance on large, centralized AI providers is diminishing.

Open-source models can be downloaded and modified in ways their creators can't control or restrict, and many are small enough and capable enough to be run locally.

Stripping the guardrails from an open-source AI model, a technique called ablation, requires specialist knowledge. This process removes the refusal behavior from a free, open-source AI model without requiring retraining, and results in a model that will answer virtually any request. Criminals don't need that expertise themselves. Thousands of models labeled "ablated," "uncensored," or similar are freely available for

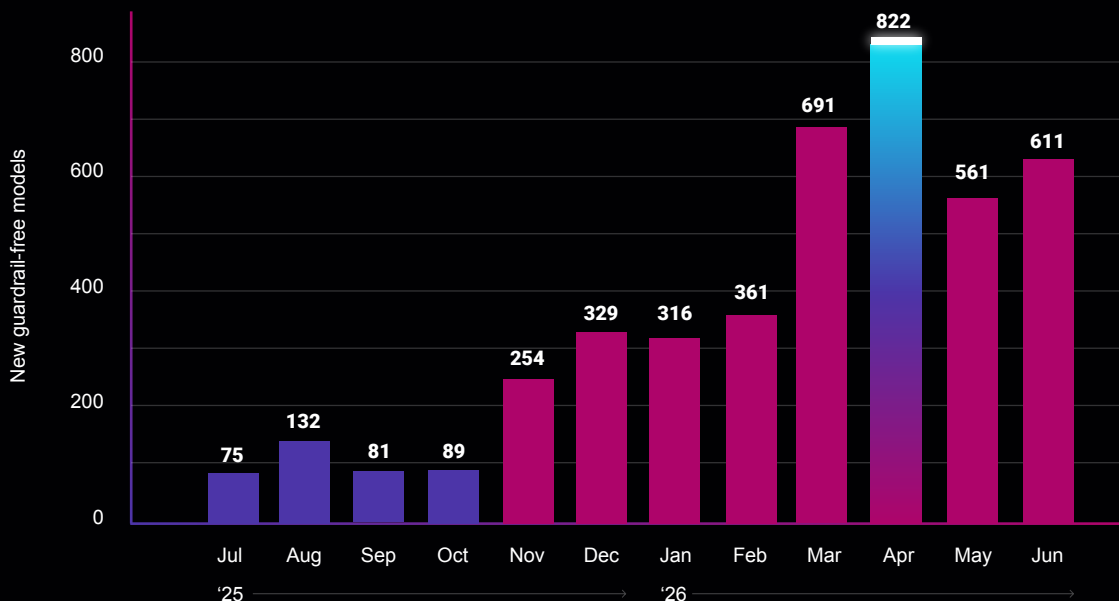
download on platforms like Hugging Face. No technical skill required.

In July 2026, our researchers found over 6,000 models published openly on Hugging Face under self-declared guardrail-free labels: "ablated," "uncensored," "heretic," "decensored," and "unfiltered." Together, they accumulated tens of millions of downloads in a single 30-day window. Publishing volume increased around 3-fold between the final quarter of 2025 and the second quarter of 2026.

A criminal weighing GhostGPT at \$50 a week against a free model that runs offline on their own machine, with no provider logging the prompts, has an easy decision.

### Hundreds of new guardrail-free models appear every month

New guardrail-free models on Hugging Face, by month first published





## Insider threats

Claude Code, Anthropic's agentic coding assistant, became one of the defining productivity stories of late 2025. Then, in January 2026, OpenClaw arrived.

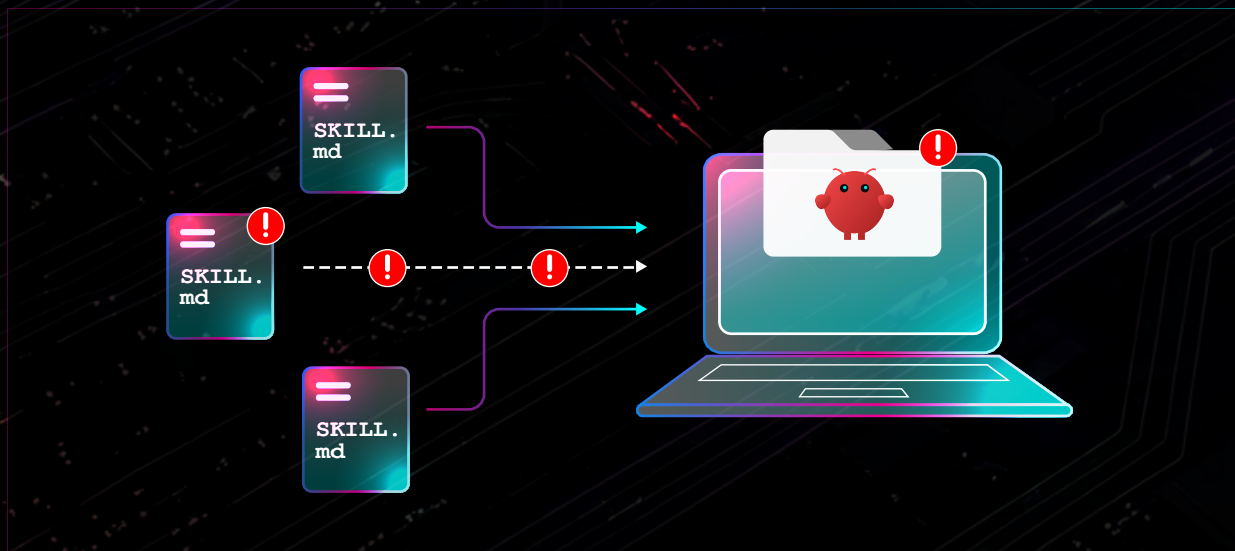
OpenClaw is a personal AI agent that runs on your computer with broad access to your files, credentials, and connected systems. It went viral, exploding in popularity and triggering a Mac mini shortage as users rushed to buy hardware to run it.<sup>8</sup> For millions, it was their first experience with an AI agent they could run themselves, locally, continuously. The enthusiasm was immediate and enormous.

So was the interest from criminals.

## The ghost in the machine

Skills are plugins that extend an AI agent's capabilities the way apps extend a smartphone. Install a skill, and your agent can act like an expert web designer, open Word documents, book meetings, or analyze a specific type of data in a predictable, repeatable way.

Skills can be discovered and installed from public marketplaces in seconds, exercise considerable influence over agents that are highly privileged, and often operate with few if any mandatory security controls. It was perhaps inevitable therefore that OpenClaw's meteoric rise would be accompanied almost immediately by the emergence of an entirely new type of threat: malicious agent skills.





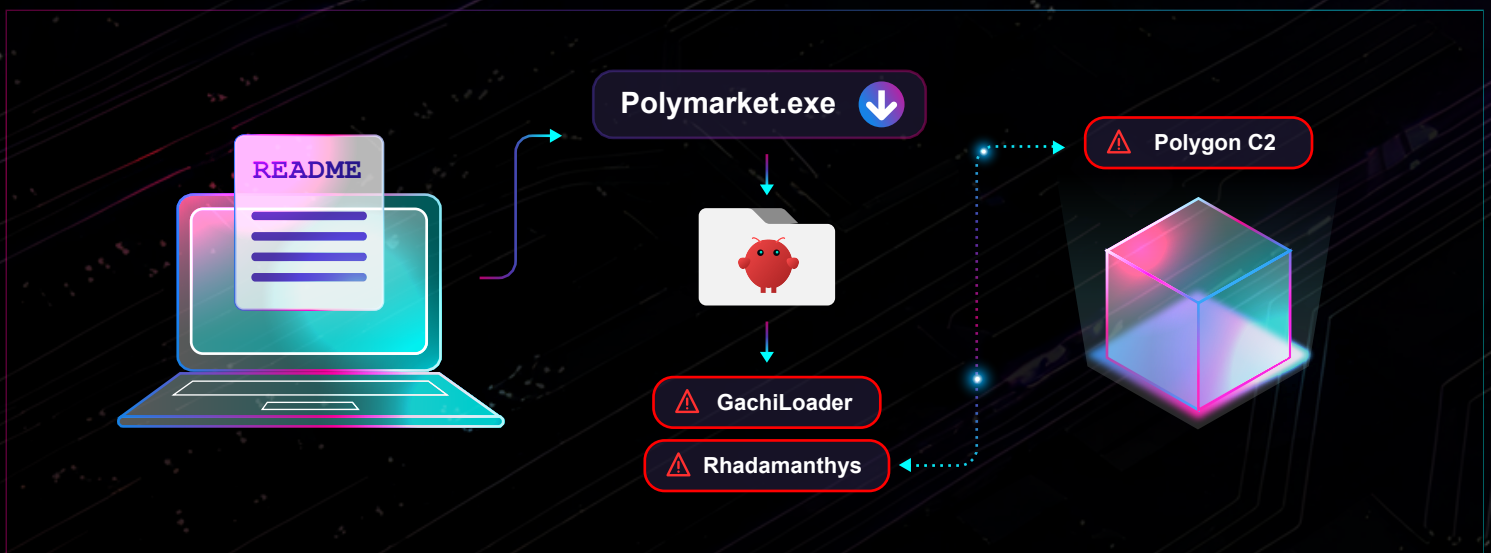
The first malicious skills were discovered on OpenClaw's official marketplace, ClawHub, within days of its launch, and within a month, security researcher @chiefofautism claimed to have found over 1,000, including the site's #1 most downloaded skill.<sup>9</sup>

An analysis by our own researchers documented malicious skills that stole credentials and exfiltrated data, installed malware, and even poisoned an agent's core memory files to ensure persistence.<sup>10</sup> Malicious behavior was typically included in otherwise functional skills, so that agents operated as expected while executing harmful operations in the background.

In April 2026, with security scanners beginning to search for harmful payloads in agent skills, criminals pivoted again, and our researchers discovered a new evolution of the malicious skill hiding on ClawHub.

The skill itself contained no malicious code. Instead, it contained a **README** file, a locally saved Telegraph blog post titled *How Building a Weather Polymarket Bot with OpenClaw Skill and turn 100\$ → 8000\$*, which explains how an agent can exploit differences between NOAA weather forecasts and Polymarket's temperature-based betting markets. It was a fabrication, constructed to look like the kind of get-rich-quick-with-AI content flooding the internet in early 2026.

The **README** instructed users to download a binary called `Polymarket.exe` and run OpenClaw `Polymarket.exe` to complete the setup.





Users who downloaded the binary would receive either a Node.js single executable application (SEA) or an Electron application. In either case, the download delivered the latest iteration of GachiLoader, which used fileless injection to install Rhadamanthys, a sophisticated infostealer that harvests credentials that can be used in identity-based attacks.

GachiLoader's command-and-control infrastructure was hosted on a Polygon blockchain smart contract, making it immutable and impossible to take down through conventional means.

The GitHub account hosting the payload, blueberrywoodsym, was purpose-built for OpenClaw targets, with its three releases named to match the stages a target would expect to encounter during a legitimate skill installation.

Detection rate across major endpoint security products at the time of discovery: near zero. **ThreatDown Endpoint Protection caught it.<sup>11</sup>**

Since the rapid adoption of AI agents in early 2026, two distinct classes of malicious skills have emerged. In the first, the skill itself is malicious, manipulating the agent directly. In the second, the skill is a lure that manipulates the agent's human operator rather than the agent itself, using detection-resistant social engineering tricks.

What both attack classes share is the same vulnerability: the enthusiasm people have for AI tools. Criminals aren't just using AI. They're exploiting the trust and excitement that AI has generated, turning the rush to adopt these tools into an attack surface.

What both attack classes share is the same vulnerability:  
the enthusiasm people have for AI tools.



## Shadow AI

**\$670K**

premium on breaches  
involving shadow AI

**100+**

separate OpenClaw  
security advisories

**1 in 5**

organizations suffered a breach  
linked to shadow AI in 2025

As organizations struggle to embrace, govern, and monitor AI, employees are increasingly adopting AI tools without telling IT: uploading sensitive documents, entering credentials, and connecting company systems to platforms that haven't been vetted or secured. Nearly half of employees who use generative AI at work do so through personal, unmanaged accounts.<sup>12</sup>

The financial consequences are already being felt. One in five organizations suffered a breach linked to shadow AI in 2025, costing an average of \$670,000 more than a standard incident.<sup>13</sup>

The scale of the problem is visible in the tools employees are reaching for. By the end of March 2026, approximately half a million OpenClaw instances were exposed on the public internet.<sup>14</sup> Most had been installed by employees without their IT team's knowledge, creating an invisible attack surface rife with opportunities for exploitation. Between February and April 2026, security researcher Jerry Gamblin's public OpenClaw vulnerability tracker logged well over 100 separate security advisories.<sup>15</sup>

AI tools have also become high-value targets in their own right, as they not only give access to conversations and the data shared within them, but also unlock connected systems, stored files, and integrated services.<sup>16</sup> The passwords, OAuth credentials, API keys, and session tokens used to access AI tools have joined the ever-expanding list of secrets you must defend against identity attacks.

Every AI tool account an employee creates without IT's knowledge is a potential identity attack waiting to happen.





# Cybercrime tomorrow

The era in which software vulnerabilities go unexploited because finding them is hard is **coming to an end.**



# Cybercrime tomorrow, six months to prepare

In May 2026, Google's Threat Intelligence Group identified what it believed to be the first zero-day exploit developed by criminals using AI. Only a month earlier, Anthropic had unveiled a general-purpose AI model so good at finding software vulnerabilities that it withheld it from public release.

The two discoveries signal that the era in which software vulnerabilities go unexploited because finding them is hard is coming to an end. What comes next will test every organization.

## The exploit window closes

Software exploits are a cornerstone of cybercrime, used by criminals to perform malicious tasks like breaching servers, elevating privileges, and extracting data.

Where they can, threat actors exploit zero-day vulnerabilities that are unknown to defenders. Where they can't, they reverse engineer newly released vendor patches, with the aim of creating exploits they can use in the weeks or months that elapse between defenders becoming aware that a patch exists and actually applying it.



## The first AI-built zero-day

Last year, landmark vulnerability discoveries by Google's Big Sleep,<sup>17</sup> and the arrival of bug bounty hunting AI XBOW at the top of the global HackerOne leaderboard, demonstrated that it was possible for AI to outperform humans at finding security flaws in software.<sup>18</sup> Both projects were designed to strengthen security rather than damage it, but the capabilities they demonstrated were so obviously useful to attackers that it was only a matter of time before criminals acquired them.

We didn't have long to wait. In May 2026, Google's Threat Intelligence Group identified what it believes to be the first known zero-day exploit developed by criminals using AI. The

target was a popular open-source, web-based system administration tool. The exploit, implemented as a Python script, was designed to bypass two-factor authentication and enable identity-based attacks.

Although Google disrupted the campaign before mass exploitation could begin, the significance of the finding wasn't the exploit itself, but what it confirmed: AI is now being used by criminal groups to discover and weaponize vulnerabilities that didn't exist in any public database.

Alarmingly, just as the world discovered that criminals were actively using AI to find zero-days, a frontier AI model appeared that foreshadows a quantum leap in their ability to do so.<sup>19</sup>

Alarmingly, just as the world discovered that criminals are actively using AI to find zero-days, a frontier AI model appeared that foreshadows a quantum leap in their ability to do so.



## Mythos

In April 2026, Anthropic unveiled its latest frontier AI, Claude Mythos, and simultaneously withheld it from public release.<sup>20</sup>

The general-purpose AI model had surprised its creator by proving far more adept at exploiting software than anyone had anticipated. Its creator warned it “could reshape cybersecurity” if it were released into the hands of criminals.

Anthropic reported that during testing, Mythos Preview found and exploited zero-day vulnerabilities in every major operating system and web browser it tested, including uncovering a 27-year-old bug in the obsessively secure OpenBSD. Engineers with no formal security background were able to use it to find remote code execution vulnerabilities while they slept.<sup>21</sup>

Mythos also proved capable of constructing sophisticated multi-stage exploits and successfully carrying out complex attack chains that can take human professionals days of work. In testing by the UK AI Security Institute, Mythos became the first model to complete a 32-step simulated corporate-network takeover end-to-end, a simulation estimated to require human professionals approximately 20 hours to complete.<sup>22</sup>

In response, Anthropic gave a select group of major technology companies and infrastructure providers exclusive access to Mythos Preview through a controlled program called Project Glasswing, so defenders could secure critical software before Mythos-like models reached anyone else. Claims that the move was a publicity stunt were quickly silenced. After applying an early version of



Mythos to its codebase, Firefox shipped a staggering 423 bug fixes in April 2026, compared to 31 exactly a year earlier.<sup>23</sup> Mozilla's CTO described the experience as giving the team "vertigo."<sup>24</sup>

Seemingly in response to Mythos, on June 2, the US government issued an executive order directing the National Security Agency (NSA) to develop a classified benchmarking process to assess the cybersecurity capabilities of frontier AI models, introducing a new "covered frontier model" designation whose threshold will be set through that process, and creating a mechanism for the government to access those models for up to 30 days before releasing them to other trusted partners. It also established an AI cybersecurity clearinghouse to coordinate vulnerability scanning and patch distribution across critical infrastructure and directed the Cybersecurity and Infrastructure Security Agency (CISA) to expand federal programs providing AI-enabled defensive tools.<sup>25</sup> It was

Its creator warned it **"could reshape cybersecurity"** if it were released into the hands of criminals.

the clearest possible signal that cybersecurity had changed forever.

On June 9, Anthropic launched Fable 5, a Mythos-class model "made safe for general use" and Mythos 5, a less restricted model that would only be available to cyberdefenders and infrastructure providers.<sup>26</sup> Three days later, the US government issued an export control directive suspending access to both models for any foreign national, inside or outside the United States, after discovering a jailbreak for Fable 5. The export restriction was broad enough that Anthropic disabled both models for all customers to ensure compliance.

But while Fable 5 and Mythos 5 were unavailable, AI development marched on. On June 22, OpenAI expanded its answer to Project Glasswing, the similarly restricted Daybreak program, and launched GPT-5.5-Cyber, a frontier model optimized for vulnerability discovery and exploitation.<sup>27</sup> Others will soon follow. Anthropic's CEO assessed that models from China with comparable capability to Mythos are likely to be broadly available within 6 to 12 months.<sup>28</sup>

The suspension lasted until June 30, when the export controls were lifted. Fable 5 returned to general availability on July 1. Mythos 5 remains available only to vetted cyberdefenders and infrastructure providers, per Anthropic's original design, not government restriction.<sup>29</sup>



It's only a matter of time, perhaps as little as a few months, before the criminal ecosystem that already sells access to jailbroken frontier models and guardrail-free open-source ones is offering Mythos-class AI.

In the near term though, Mythos-class models are beginning to reach the hands of defenders, which sounds like good news. However, it isn't straightforward. Models capable of finding vulnerabilities at this scale will generate patch volumes that dwarf anything organizations currently manage. Because every new patch is an opportunity for criminals to reverse engineer an exploit, the number of exploits available to criminals in the short term could grow massively.

Simultaneously, organizations that struggle with their current patching backlog won't suddenly find ten times the capacity to handle ten times the fixes. They will fall behind, and the window in which exploits remain effective will grow.

The volume of work that Mythos-class models will create for defenders is as significant as the threat they pose in the wrong hands.

#### Firefox bug fixes shipped in April, year over year<sup>23</sup>

April 2025

31 fixes

April 2026

423 fixes

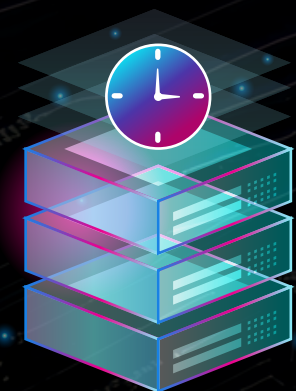
The volume of work that Mythos-class models will create for defenders is as significant as the threat they pose in the wrong hands.



## Conclusion

Everything in this report is something that is tangible and real. There are no “what ifs”. The attack on the water utility in Monterrey was real. The marketplace for criminal AI, the first AI-built zero-day, the tens of millions of downloads of guardrail-free AI models, and the poisoned skill on the OpenClaw marketplace are real.

The question you face isn’t whether AI-native attacks are a threat you’ll face. It’s whether you’ll be prepared for them when they do. The transition to a cybercrime ecosystem shaped by AI has begun, and there are perhaps six months to prepare, across three key areas:



### Patch like the clock is running.

Mythos and its successors could inundate organizations with a huge volume of extra software updates. Organizations that stay on top of patching with automation and effective vulnerability assessment and patch management will be more secure. Those that don’t will face an ever-growing backlog of unpatched systems, longer periods unprotected, and criminals armed with countless new exploits.



### Monitor continuously, 24/7.

AI makes every threat actor faster, smarter, and more scalable, but it doesn’t create new tactics. Data held on endpoints is still the target and stolen identities remain the dominant form of access. Organizations must set themselves up to react to the early signs of an AI-driven attack rapidly, with a security operations center (SOC) or a managed detection and response (MDR) service.





### **Find and govern your shadow AI.**

Unsanctioned AI tools, agent skills, and MCP connections are a growing compliance, privacy, and security risk your IT team can't see. Use AI detection and response to discover the shadow AI attack surface inside your organization.

The future of cybercrime is already here. It's simply not yet evenly distributed. The organizations that come through this moment intact will be the ones that saw it clearly and acted while the clock still favored the prepared.



1. ThreatDown (2025), AI has “fully defeated” most of the ways people authenticate, <https://www.threatdown.com/blog/ai-has-fully-defeated-how-most-people-authenticate/>
2. Hoxhunt (2025), AI-Powered Phishing Outperforms Elite Red Teams in 2025, <https://hoxhunt.com/blog/ai-powered-phishing-vs-humans>
3. Dragos (2026), AI in the Breach: How an Adversary Leveraged AI to Target a Water Utility's OT, <https://www.dragos.com/blog/ai-assisted-ics-attack-water-utility>
4. Anthropic (2025), Threat Intelligence Report: August 2025, <https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf>
5. Anthropic (2025), Disrupting the first reported AI-orchestrated cyber espionage campaign, <https://www-cdn.anthropic.com/d7dd50dd1185f59be051b307150d877f2b82bd2c.pdf>
6. Google Cloud (2025), GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools, <https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools>
7. Google Cloud (2025), GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools, <https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools>
8. PBX Science (2026), OpenClaw: GitHub's Fastest-Ever Rising Star Become's 2026's First Major AI Security Disaster, <https://pbxscience.com/openclaw-githubs-fastest-ever-rising-star-becomes-2026s-first-major-ai-security-disaster/>
9. @chiefofautism (2026), The #1 most downloaded skill on OpenClaw marketplace was MALWARE, <https://x.com/chiefofautism/status/2024483631067021348>
10. ThreatDown (2026), Weaponizing autonomy: The rise of malicious AI agent skills, <https://www.threatdown.com/blog/weaponizing-autonomy-the-rise-of-malicious-ai-agent-skills/>
11. ThreatDown (2026), GachiLoader adopts AI skill lure, <https://www.threatdown.com/blog/gachiloader-adopts-ai-skill-lure-from-fake-openclaw-readme-to-rhadamanthys-infostealer/>
12. Netskope (2026), Netskope Cloud and Threat Report 2026, <https://www.netskope.com/resources/cloud-and-threat-reports/cloud-and-threat-report-2026>
13. IBM (2025), Cost of a Data Breach 2025, <https://www.ibm.com/reports/data-breach>
14. VentureBeat (2026), OpenClaw has 500,000 instances and no enterprise kill switch, <https://venturebeat.com/security/openclaw-500000-instances-no-enterprise-kill-switch>
15. Gamblin, J. (2026), OpenClaw CVEs, <https://github.com/jgamblin/OpenClawCVEs>
16. IBM (2026), X-Force Threat Intelligence Index 2026, <https://www.ibm.com/reports/threat-intelligence>
17. Google Cloud (2025), Cloud CISO Perspectives: Our Big Sleep agent makes a big leap, <https://cloud.google.com/blog/products/identity-security/cloud-ciso-perspectives-our-big-sleep-agent-makes-big-leap>
18. XBOW (2025), XBOW on HackerOne: What's Next, <https://xbow.com/blog/xbow-on-hackerone-whats-next>
19. Google Threat Intelligence Group (2026), GTIG AI Threat Tracker: Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access, <https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access>
20. Anthropic (2026), Project Glasswing, <https://www.anthropic.com/project/glasswing>
21. Anthropic Red Team Blog (2026), Assessing Claude Mythos Preview's cybersecurity capabilities, <https://red.anthropic.com/2026/mythos-preview/>
22. AISI (2026), Our evaluation of Claude Mythos Preview's cyber capabilities, <https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities>
23. TechCrunch (2026), How Anthropic's Mythos has rewritten Firefox's approach to cybersecurity, <https://techcrunch.com/2026/05/07/how-anthropics-mythos-has-rewritten-firefoxs-approach-to-cybersecurity/>
24. Mozilla (2026), The zero-days are numbered, <https://blog.mozilla.org/en/privacy-security/ai-security-zero-day-vulnerabilities/>
25. The White House (2026), Promoting Advanced Artificial Intelligence innovation and security, <https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/>
26. Anthropic (2026), Claude Fable 5 and Claude Mythos 5, <https://www.anthropic.com/news/claude-fable-5-mythos-5>
27. OpenAI (2026), Daybreak: Tools for securing every organization in the world, <https://openai.com/index/daybreak-securing-the-world/>
28. CNBC (2026), Anthropic CEO warns of cyber 'moment of danger' as AI exposes thousands of vulnerabilities, <https://www.cnbc.com/2026/05/05/anthropic-ceo-cyber-moment-of-danger-mythos-vulnerabilities.html>
29. Anthropic (2026), Redeploying Claude Fable 5, <https://www.anthropic.com/news/redeploying-fable-5>