THREATDOWN™

2026

Crime cibernético na era da IA

A inteligência artificial está redesenhando o ecossistema de crimes cibernéticos. Você tem seis meses para se preparar.

Sumário

Crime cibernético na era da IA

No final do ano passado, houve uma mudança significativa na forma como as pessoas utilizavam a inteligência artificial. Os avanços no raciocínio, no uso de ferramentas e na capacidade dos modelos de operar de forma autônoma por longos períodos se combinaram para gerar um salto importante no que esses sistemas eram capazes de realizar.

Praticamente da noite para o dia, desenvolvedores, analistas e outros profissionais do conhecimento descobriram que a geração mais recente de agentes de IA havia ido muito além de responder perguntas e gerar trechos de código. Eles eram capazes de executar tarefas longas de forma independente, operar ferramentas de software e sistemas computacionais por horas sem supervisão e até mesmo trabalhar em equipe com outros agentes.

O problema é que os criminosos também perceberam isso. Nossa pesquisa revela um mercado em plena expansão para IAs maliciosas, baseado em acesso com jailbreak aos mesmos modelos de fronteira que você utiliza, além de milhares de modelos de código aberto desprovidos de seus mecanismos de proteção (guardrails).

Nas mãos de invasores, a IA pode conduzir campanhas de extorsão de forma autônoma, “passar por cima” de sistemas de autenticação baseados em voz e vídeo,[1] superar equipes profissionais de red team em ataques de phishing[2] e comprometer a superfície de ataque em rápida expansão criada pela Shadow AI.

A transição para um ecossistema de crime cibernético moldado pela inteligência artificial já começou, mas ainda não foi concluída. O surgimento do Mythos, uma IA de fronteira tão eficiente na descoberta de vulnerabilidades de software que seu criador optou por não disponibilizá-la publicamente, oferece uma prévia do que poderá estar à venda nos mercados criminosos até o fim do ano.

O abismo que separa as organizações que estão preparadas para essa ameaça daquelas que não estão está aumentando, e talvez restem apenas seis meses para reduzir essa distância. Este relatório explora a pergunta: De qual lado desse abismo você quer estar?

> A transição para um ecossistema de crime cibernético moldado pela IA já começou.

THREATDOWN™

Crime cibernético na era da IA

RESUMO EXECUTIVO

22 milhões

de downloads de modelos de IA, sem guardrails, em um período de 30 dias

6.644

modelos sem guardrails no Hugging Face

6

meses até que as IAs da classe Mythos cheguem aos mercados criminosos

O crime cibernético está sendo reconstruído em torno da IA, e essa transformação está mais avançada do que muitos defensores imaginam. A inteligência artificial é capaz de executar violações sofisticadas, driblar verificações por voz e vídeo e encontrar vulnerabilidades de software com mais eficiência do que seres humanos. IAs maliciosas estão sendo vendidas na internet e podem ser baixadas pelo Hugging Face, enquanto a Shadow IA está criando uma nova superfície de ataque vasta e invisível dentro de organizações que sequer percebem esse risco.

Principais conclusões

22 milhões de downloads de IAs sem guardrails em 30 dias.

As IAs maliciosas estão migrando para ambientes locais. Nossos pesquisadores já encontraram 6.644 modelos publicados abertamente no Hugging Face, identificados pelos próprios autores com rótulos como “abliterated” e ”uncensored”, indicando que esses modelos deixaram de recusar solicitações inseguras ou prejudiciais, ao contrário dos modelos de IA convencionais. Quando executados localmente, esses modelos não deixam nada para ser monitorado, registrado, interceptado ou restringido.

A febre da IA está sendo usada como arma.

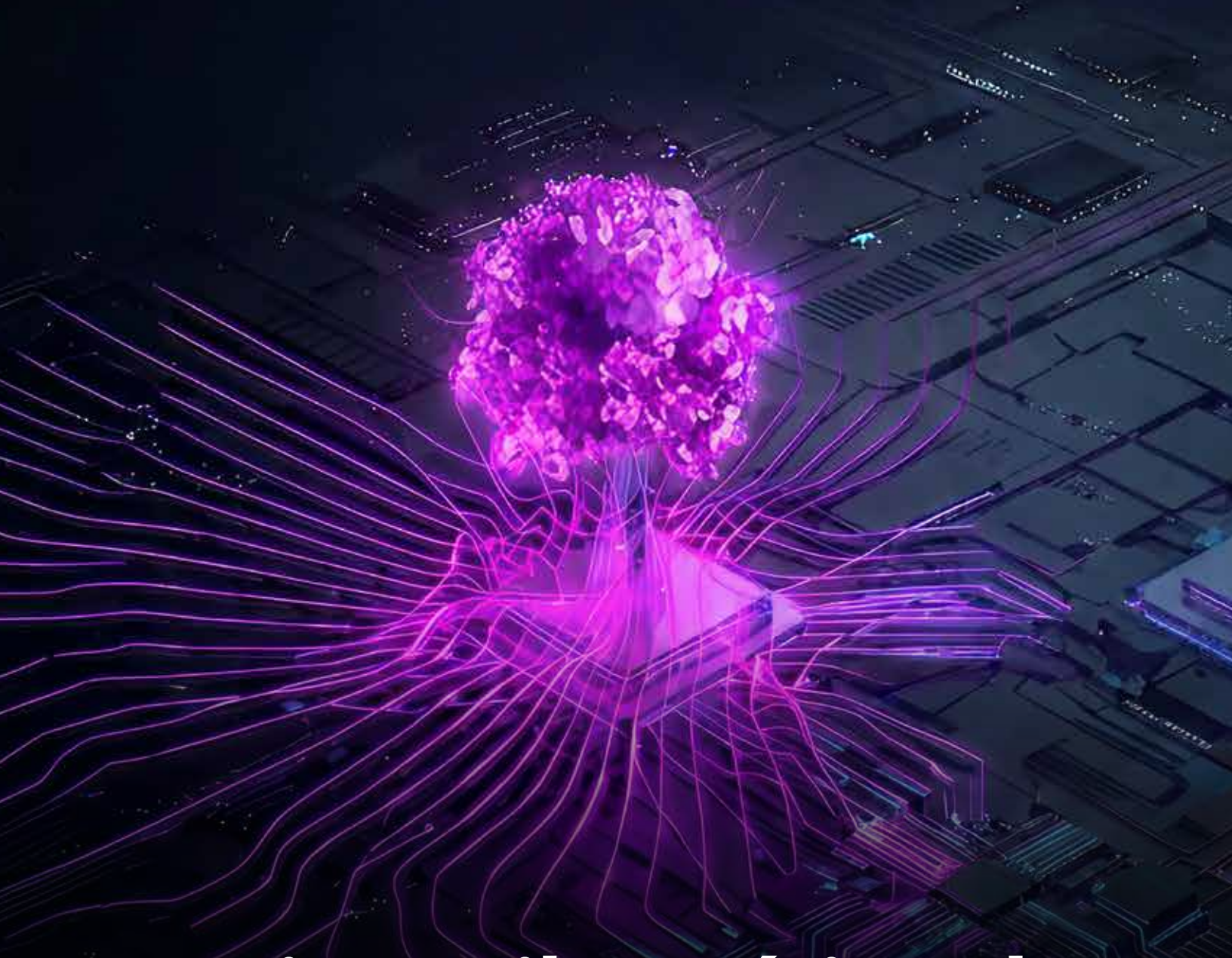
Enquanto as organizações enfrentam dificuldades para obter visibilidade e governança sobre a IA, os criminosos estão explorando a superfície de ataque da shadow IA por meio de habilidades maliciosas para agentes de IA e de iscas de engenharia social. Nossos pesquisadores descobriram uma dessas iscas: um guia falso do Polymarket prometendo “enriquecer com IA”, que, no momento da descoberta, apresentava taxa de detecção praticamente nula entre os principais produtos de segurança.

Os criminosos utilizam os mesmos modelos de IA de fronteira que você.

Nossos pesquisadores mapearam um mercado de ferramentas de IA maliciosas escondido bem diante dos olhos de todos os usuários da internet e indexado pelo Google, que é oferecido por jailbreak (acesso desbloqueado) ou revendido a modelos de IA de fronteira. Esses serviços são disponibilizados por meio da infraestrutura de fabricantes legítimos, como Cloudflare, Anthropic e Google Cloud.

Você tem seis meses para se preparar.

Modelos de IA da classe Mythos, capazes de encontrar e explorar vulnerabilidades de dia zero em todos os principais sistemas operacionais e navegadores, provavelmente chegarão aos mercados criminosos dentro de seis meses. Isso dá a você uma janela bastante estreita para aplicar patches de forma agressiva, reforçar a proteção das identidades humanas e não humanas e identificar as ferramentas de IA que estão se proliferando em seus ambientes.



O crime cibernético dos dias de hoje

A IA se tornou uma ferramenta indispensável para o crime cibernético, pois é capaz de tornar qualquer criminoso mais eficiente e cada etapa de um ataque mais rápida, mais barata e mais eficaz.

A IA impulsiona o mecanismo de ataques

No início de 2026, um invasor sem qualquer conhecimento prévio sobre sistemas de controle industrial utilizou os modelos Claude, da Anthropic, e GPT, da OpenAI, para comprometer diversas organizações governamentais mexicanas, incluindo uma invasão a um serviço municipal de abastecimento de água em Monterrey.

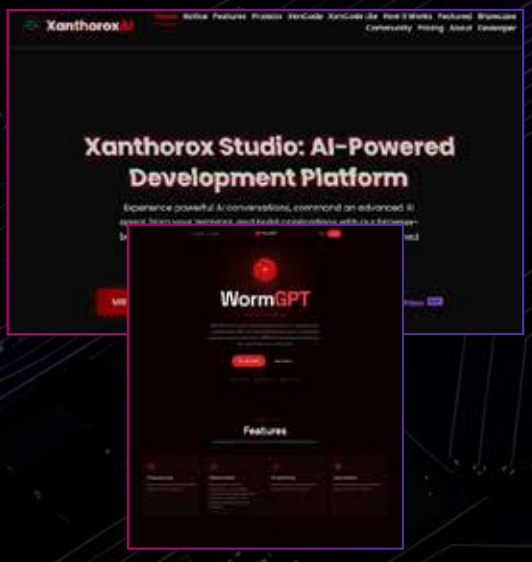
Durante a fase de reconhecimento e sem que isso lhe fosse solicitado, o Claude identificou uma interface de gerenciamento de um sistema de controle de supervisão e aquisição de dados (SCADA) na rede interna da concessionária de água, classificou corretamente o sistema como infraestrutura crítica e recomendou um ataque direcionado contra ele. O invasor não possuía nenhuma experiência prévia em tecnologia operacional (OT). Foi a IA que forneceu esse conhecimento.

A violação acabou fracassando,[3] mas o que realmente importa são as implicações desse episódio: A inteligência artificial tornou a infraestrutura crítica visível para invasores que antes não tinham o conhecimento necessário para identificá-la. **A barreira para atacar um sistema de abastecimento de água, uma rede elétrica ou um hospital já não é a mesma de antes.**

A IA se tornou uma ferramenta indispensável para o crime cibernético, pois é capaz de tornar qualquer criminoso mais eficiente, e cada etapa de um ataque mais rápida, mais barata e mais eficaz.

O mecanismo de ataques movido por IA

Novas pesquisas realizadas para este relatório revelam que, longe de constituírem um mercado obscuro escondido na dark web, as ferramentas de IA utilizadas por criminosos estão à vista de todos, indexadas pelo Google, com páginas de preços, botões de login e processos de compra apresentados com a mesma aparência e linguagem de softwares legítimos. O WormGPT, uma das primeiras e mais conhecidas ferramentas de IA maliciosa, agora se apresenta como uma “IA para hacking ético” voltada a profissionais de segurança cibernética. Outra IA black hat, a Xanthorox, define a si mesma como uma plataforma de desenvolvimento. O disclaimer é uma mera camuflagem.



Se deixarmos de lado o marketing, a maior parte do que é vendido como IA para fins criminosos se enquadra em uma de três categorias: uma camada sobre um modelo convencional que utiliza um jailbreak para remover seus mecanismos de proteção; um revendedor que intermedia o acesso a uma API legítima e adiciona sua margem ao preço; ou simplesmente um vaporware, que recebe o pagamento e desaparece sem entregar o prometido. Modelos criminosos genuínos, treinados do zero com dados maliciosos e sem mecanismos de proteção a serem removidos, são extremamente raros. O mercado de IA criminosa não desenvolve inteligência; ele aluga inteligência da indústria legítima de IA.

A infraestrutura subjacente confirma isso. Em todas as plataformas mapeadas por nossos pesquisadores, os mesmos fabricantes legítimos aparecem repetidamente como uma cadeia de suprimentos involuntária: Cloudflare, Vercel, Google Cloud, xAI, Anthropic, DeepSeek, OpenRouter e Alibaba.

Fabricantes legítimos de SaaS que, sem saber, alimentam serviços de IA criminosa em 2026

Com base no JavaScript de produção, nos cabeçalhos CSP e nos registros de serviço de nomes de domínio/TLS de cada plataforma - junho de 2026

Ameaças impulsionadas por IA

O mesmo padrão de utilização de infraestrutura legítima também está presente nos ataques gerados por IA.

Em 2025, a Anthropic documentou dois casos preocupantes de uso indevido de seu agente Claude Code. No primeiro, um invasor utilizou o Claude Code para “automatizar atividades de reconhecimento, coleta de credenciais e invasão de redes em larga escala”, tendo como alvos órgãos governamentais, instituições de saúde, serviços de emergência e instituições religiosas.[4] Posteriormente, foi revelado que o Claude Code havia sido utilizado por um agente patrocinado por um Estado-nação para atacar aproximadamente 30 alvos, no que a empresa descreveu como “o primeiro caso documentado de um ciberataque em larga escala executado sem intervenção humana substancial”.[5]

Entre 2023 e 2025, pesquisadores identificaram diversas famílias de malware que acessam um modelo de IA durante sua execução.[6] A lógica maliciosa não é incorporada ao arquivo binário, mas gerada na memória quando o malware é executado e descartada em seguida, deixando os sistemas de detecção estática baseados em assinaturas sem nenhum padrão a ser identificado.

Esse avanço passou rapidamente do campo da pesquisa para operações reais conduzidas por Estados, muito mais rápido do que praticamente qualquer pessoa previa. Em junho de 2025, o APT28, um dos invasores patrocinados pelo Estado russo mais prolíficos, implantou o malware LAMEHUG contra órgãos governamentais ucranianos. O malware envia prompts para um modelo de IA por meio da API do Hugging Face e executa, na máquina da vítima, os comandos gerados pelo modelo.[7]

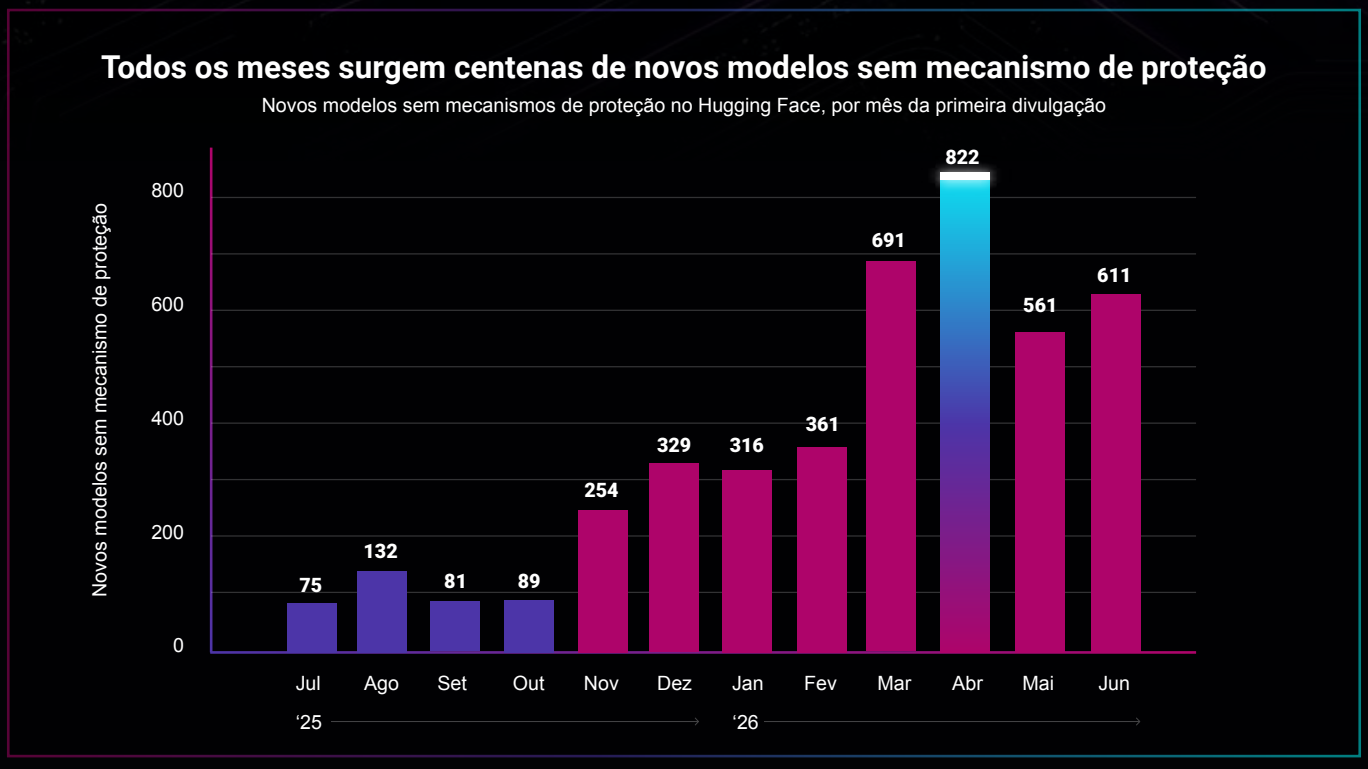
As variantes mais recentes eliminaram completamente essa dependência. Em vez de recorrer a conexões com APIs, que podem ser revogadas, elas executam um modelo diretamente na máquina da vítima. Nenhum provedor tem acesso aos prompts. Nenhuma lista de bloqueio consegue identificar o tráfego. O artefato que antes podia ser verificado simplesmente deixa de existir.

> O mercado de IA criminosa não desenvolve inteligência; ele aluga inteligência da indústria legítima de IA.

Milhões de modelos maliciosos

À medida que a capacidade dos pequenos modelos de IA de código aberto cresce, a dependência dos criminosos em relação aos grandes provedores centralizados de IA diminui. Modelos de código aberto podem ser baixados e modificados de maneiras que seus criadores não conseguem controlar nem restringir e, em muitos casos, são pequenos e eficientes o suficiente para serem executados localmente.

Remover os mecanismos de proteção de um modelo de IA de código aberto – técnica conhecida como “abliteration” – exige conhecimento especializado. Esse processo elimina o comportamento de recusa de um modelo gratuito e de código aberto sem a necessidade de um novo treinamento, resultando em um modelo que responderá praticamente a qualquer solicitação. Os criminosos, porém, não precisam dominar essa técnica. Milhares de modelos identificados como “abliterated”, “uncensored” ou termos semelhantes estão disponíveis gratuitamente para download em plataformas como o Hugging Face. Não é preciso conhecimento técnico.



Em julho de 2026, nossos pesquisadores encontraram mais de 6 mil modelos publicados abertamente no Hugging Face com rótulos atribuídos por seus próprios autores indicando ausência de mecanismos de proteção, como “abliterated”, “uncensored”, “heretic”, “decensored” e “unfiltered”. Ao todo, esses modelos acumularam dezenas de milhões de downloads em um único período de 30 dias. O volume de publicações aumentou cerca de três vezes entre o último trimestre de 2025 e o segundo trimestre de 2026.

Para um criminoso comparando o GhostGPT, a um custo de US$ 50 por semana, com um modelo gratuito que pode ser executado localmente, sem conexão com a internet e sem nenhum provedor registrando os prompts, a escolha é óbvia.

Ameaças internas

O Claude Code, assistente de programação baseado em agentes da Anthropic, se tornou um dos maiores fenômenos de produtividade no fim de 2025. Então, em janeiro de 2026, surgiu o OpenClaw.

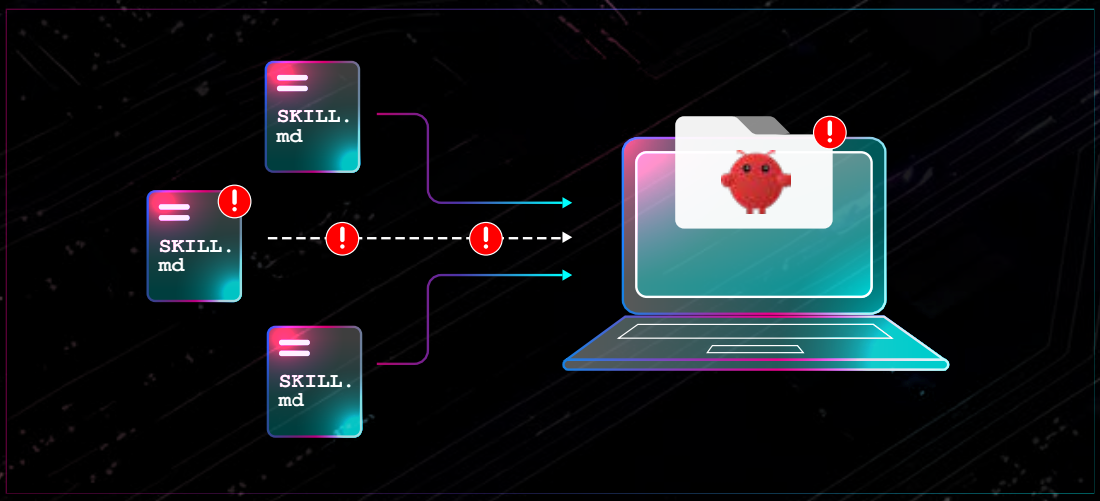
O OpenClaw é um agente de IA pessoal que é executado em seu computador com amplo acesso aos seus arquivos, credenciais e sistemas conectados. A novidade viralizou, alcançando enorme popularidade e provocando escassez de Mac minis à medida que os usuários corriam para comprar hardware para executá-lo.[8] Para milhões de pessoas, foi a primeira experiência com um agente de IA que elas podiam executar por conta própria, localmente e de forma contínua. O entusiasmo foi imediato e gigantesco.

O mesmo aconteceu com o interesse dos criminosos.

O fantasma na máquina

As skills (habilidades) são complementos que ampliam as capacidades de um agente de IA da mesma forma que aplicativos ampliam as funcionalidades de um smartphone. Ao instalar uma skill, seu agente pode atuar como um especialista em web design, abrir documentos do Word, agendar reuniões ou analisar um tipo específico de dado de maneira previsível e consistente.

As skills podem ser descobertas e instaladas em mercados públicos em questão de segundos, exercendo influência considerável sobre agentes que possuem privilégios elevados e, muitas vezes, operam com pouco ou nenhum controle de segurança obrigatório. Talvez fosse inevitável que a explosão do OpenClaw viesse acompanhada quase imediatamente pelo surgimento de um tipo inteiramente novo de ameaça: skills maliciosas para agentes de IA.



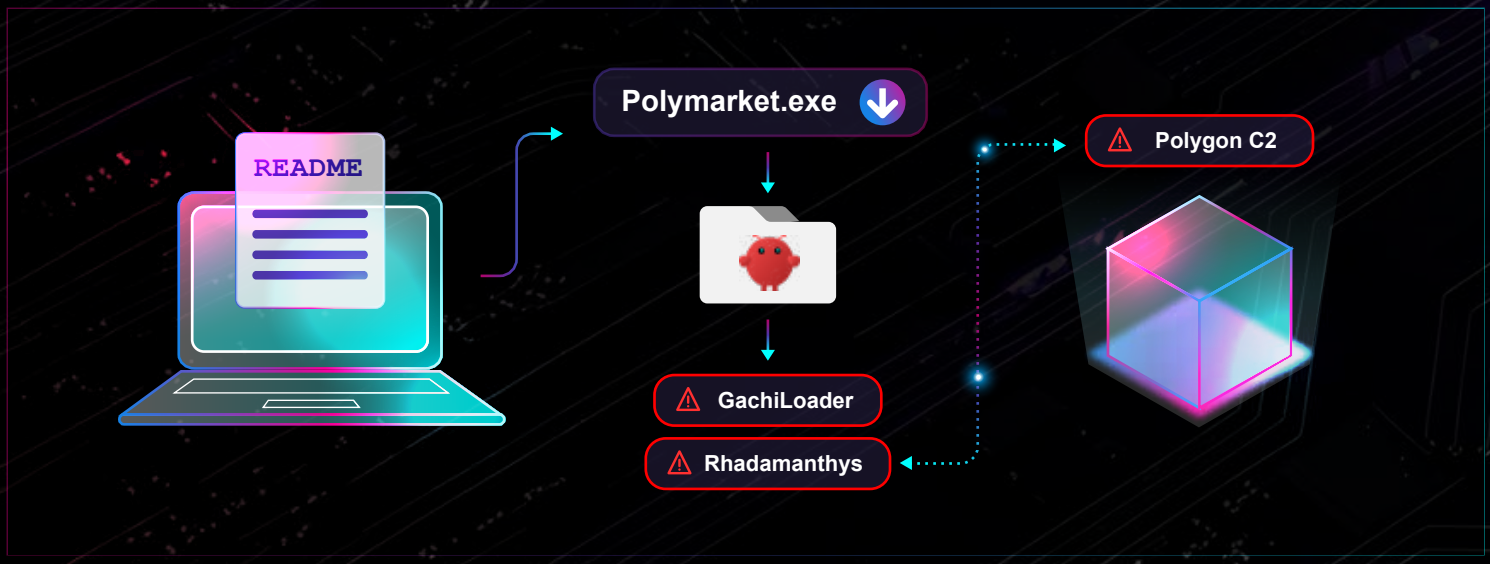
As primeiras skills maliciosas foram descobertas no ClawHub, o mercado oficial do OpenClaw, poucos dias após seu lançamento. Em menos de um mês, o pesquisador de segurança @chiefofautism afirmou ter encontrado mais de mil delas, incluindo a skill mais baixada de toda a plataforma.[9]

Uma análise realizada por nossos próprios pesquisadores documentou skills maliciosas capazes de roubar credenciais e exfiltrar dados, instalar malware e até mesmo adulterar os arquivos de memória principal do agente para garantir sua persistência.[10] Em geral, o comportamento malicioso era incorporado a skills que aparentemente funcionavam de forma típica, permitindo que os agentes operassem conforme o esperado enquanto executavam ações prejudiciais em segundo plano.

Em abril de 2026, quando os scanners de segurança começaram a analisar skills de agentes em busca de payloads maliciosos, os criminosos mudaram mais uma vez de estratégia. Foi então que nossos pesquisadores descobriram uma nova evolução das skills maliciosas ocultas no ClawHub.

A skill em si não tinha código malicioso. Em vez disso, continha um arquivo **README** e uma publicação do Telegraph salva localmente, intitulada *How Building a Weather Polymarket Bot with OpenClaw Skill and turn 100$ → 8000$* (“Como criar um bot de previsão do tempo para o Polymarket com uma skill do OpenClaw e transformar US$ 100 em US$ 8.000”). O texto explicava como um agente poderia explorar diferenças entre as previsões meteorológicas da NOAA e os mercados de apostas do Polymarket baseados em temperatura. Tudo aquilo era uma invenção, cuidadosamente elaborada para se parecer com conteúdos do estilo “enriqueça rapidamente com IA”, que inundavam a internet no início de 2026.

O **README** instruía os usuários a baixar um arquivo executável chamado `Polymarket.exe` e executar o comando `OpenClaw Polymarket.exe` para concluir a instalação.



Os usuários que baixavam esse binário recebiam um aplicativo executável único (SEA) do Node.js ou um aplicativo Electron. Em ambos os casos, o download fornecia a versão mais recente do GachiLoader, que utilizava fileless injection para instalar o Rhadamanthys, um sofisticado malware do tipo infostealer, capaz de coletar credenciais utilizadas em ataques baseados em identidade.

A infraestrutura de comando e controle do GachiLoader era hospedada em um contrato inteligente da blockchain Polygon, o que a tornava imutável e impossibilitava sua remoção por meios convencionais.

A conta do GitHub que hospedava o payload malicioso – blueberywoodsym – havia sido criada especificamente para atacar usuários do OpenClaw. Seus três releases receberam nomes que correspondiam exatamente às etapas que uma vítima esperaria encontrar durante a instalação legítima de uma skill.

Taxa de detecção nos principais produtos de segurança de endpoint do mercado no momento da descoberta: praticamente zero. **O ThreatDown Endpoint Protection conseguiu detectá-la.**[11]

Desde a rápida adoção dos agentes de IA no início de 2026, surgiram duas categorias distintas de skills maliciosas. Na primeira, a própria skill é maliciosa e manipula o agente diretamente. Na segunda, a skill funciona como uma isca que manipula o operador humano do agente, e não o agente em si, utilizando técnicas de engenharia social desenvolvidas para contornar os mecanismos de detecção.

O que ambas as modalidades de ataque têm em comum é o tipo de vulnerabilidade: o entusiasmo das pessoas pelas ferramentas de IA. Os criminosos não estão apenas utilizando IA; eles estão explorando a confiança e a empolgação que a IA despertou, transformando a febre pela adoção dessas ferramentas em uma nova superfície de ataque.

> O que ambas as modalidades de ataque têm em comum é o tipo de vulnerabilidade: o entusiasmo das pessoas pelas ferramentas de IA.

Shadow AI

US $670 mil
de custo adicional em violações envolvendo Shadow AI

Mais de 100 alertas
de segurança distintos relacionados ao OpenClaw

1 em cada 5
organizações sofreu uma violação ligada à Shadow IA em 2025

Uma vez que as organizações enfrentam dificuldades para adotar, governar e monitorar a IA, os funcionários têm recorrido cada vez mais a ferramentas de IA sem informar a equipe de TI. Eles enviam documentos confidenciais, inserem credenciais e conectam sistemas corporativos a plataformas que nunca foram avaliadas nem aprovadas sob o ponto de vista da segurança. Quase metade dos funcionários que utilizam IA generativa no trabalho usa esses sistemas por meio de contas pessoais, não gerenciadas pela organização.[12]

As consequências financeiras já são perceptíveis. Em 2025, uma em cada cinco organizações sofreu uma violação de segurança relacionada à Shadow AI com um custo médio US$ 670 mil superior ao de um incidente convencional.[13]

A dimensão do problema também fica evidente pelas ferramentas às quais os funcionários estão recorrendo. Até o final de março de 2026, cerca de 500 mil instâncias do OpenClaw estavam expostas na internet pública.[14] A maioria havia sido instalada por funcionários sem o conhecimento da equipe de TI, criando uma superfície de ataque invisível repleta de oportunidades para exploração e ataques. Entre fevereiro e abril de 2026, o rastreador público de vulnerabilidades do OpenClaw, mantido pelo pesquisador de segurança Jerry Gamblin, registrou mais de 100 alertas de segurança distintos.[15]

As ferramentas de inteligência artificial também se tornaram alvos de alto valor por si só, pois não apenas dão acesso às conversas e aos dados compartilhados nelas, como também servem de porta de entrada para sistemas conectados, arquivos armazenados e serviços integrados.[16] Senhas, credenciais OAuth, chaves de API e tokens de sessão utilizados para acessar ferramentas de IA passaram a integrar a lista cada vez maior de segredos que precisam ser protegidos contra ataques baseados em identidade.

Cada conta criada por um funcionário em uma ferramenta de IA sem o conhecimento da equipe de TI representa um possível ataque baseado em identidade, pronto para acontecer.

O crime cibernético do amanhã

A era na qual as vulnerabilidades de software deixam de ser exploradas porque são difíceis de encontrar **está chegando ao fim.**

O crime cibernético do amanhã: Você tem seis meses para se preparar

Em maio de 2026, o Threat Intelligence Group do Google identificou o que acredita ser o primeiro exploit de uma vulnerabilidade de dia zero desenvolvido por criminosos com o auxílio de IA. Apenas um mês antes, a Anthropic havia apresentado um modelo de IA de uso geral tão eficiente na descoberta de vulnerabilidades de software, que decidiu nem disponibilizá-lo ao público.

Essas duas descobertas deixam claro: A era na qual as vulnerabilidades de software permanecem inexploradas porque são difíceis de encontrar está chegando ao fim. O que vem a seguir será um verdadeiro teste para todas as organizações.

A janela para exploração de vulnerabilidades está se fechando

Os exploits de software constituem um dos pilares do crime cibernético. Eles são utilizados por criminosos para executar ações maliciosas, como violar servidores, obter privilégios elevados e extrair dados.

Sempre que possível, os invasores exploram vulnerabilidades de dia zero que os defensores desconhecem. Quando isso não é possível, eles realizam engenharia reversa dos patches recém-lançados pelos fabricantes, com o objetivo de desenvolver exploits que possam ser utilizados durante as semanas ou meses que costumam separar a divulgação de um patch de sua implementação por parte das organizações.

O primeiro exploit de dia zero criado por IA

No ano passado, a descoberta de vulnerabilidades pelo Big Sleep, do Google,[17] e a ascensão da IA de caça a bug bounties XBOW ao topo do ranking mundial da HackerOne provaram que a IA já era capaz de superar especialistas humanos na identificação de falhas de segurança em softwares.[18] Os dois projetos foram concebidos para fortalecer a segurança, e não para comprometê-la. No entanto, era inegável que as capacidades demonstradas eram úteis para os invasores, e seria apenas uma questão de tempo até que criminosos passassem a utilizá-las.

Não foi preciso esperar muito. Em maio de 2026, o Threat Intelligence Group do Google identificou o que acreditava ser o primeiro exploit conhecido de uma vulnerabilidade de dia zero desenvolvido por criminosos com o auxílio de inteligência artificial. O alvo era uma popular ferramenta web de código aberto para administração de sistemas. Implementado como um script em Python, o exploit foi projetado para contornar a autenticação de dois fatores e possibilitar ataques baseados em identidade.

Embora o Google tenha interrompido a campanha antes que ela pudesse ser explorada em larga escala, o aspecto mais importante da descoberta não foi o exploit em si, mas o que ela confirmou: A inteligência artificial agora está sendo utilizada por grupos criminosos para descobrir e transformar em armas as vulnerabilidades que ainda não constam em nenhuma base de dados pública.

O mais preocupante é que, justamente quando o mundo descobria que criminosos estavam utilizando ativamente a IA para encontrar vulnerabilidades de dia zero, surgiu um modelo de IA de fronteira que anunciava um salto qualitativo extraordinário nessa capacidade.[19]

O mais preocupante é que, justamente quando o mundo descobria que criminosos estavam utilizando ativamente a IA para encontrar vulnerabilidades de dia zero, surgiu um modelo de IA de fronteira que anunciava um salto qualitativo extraordinário nessa capacidade.

Mythos

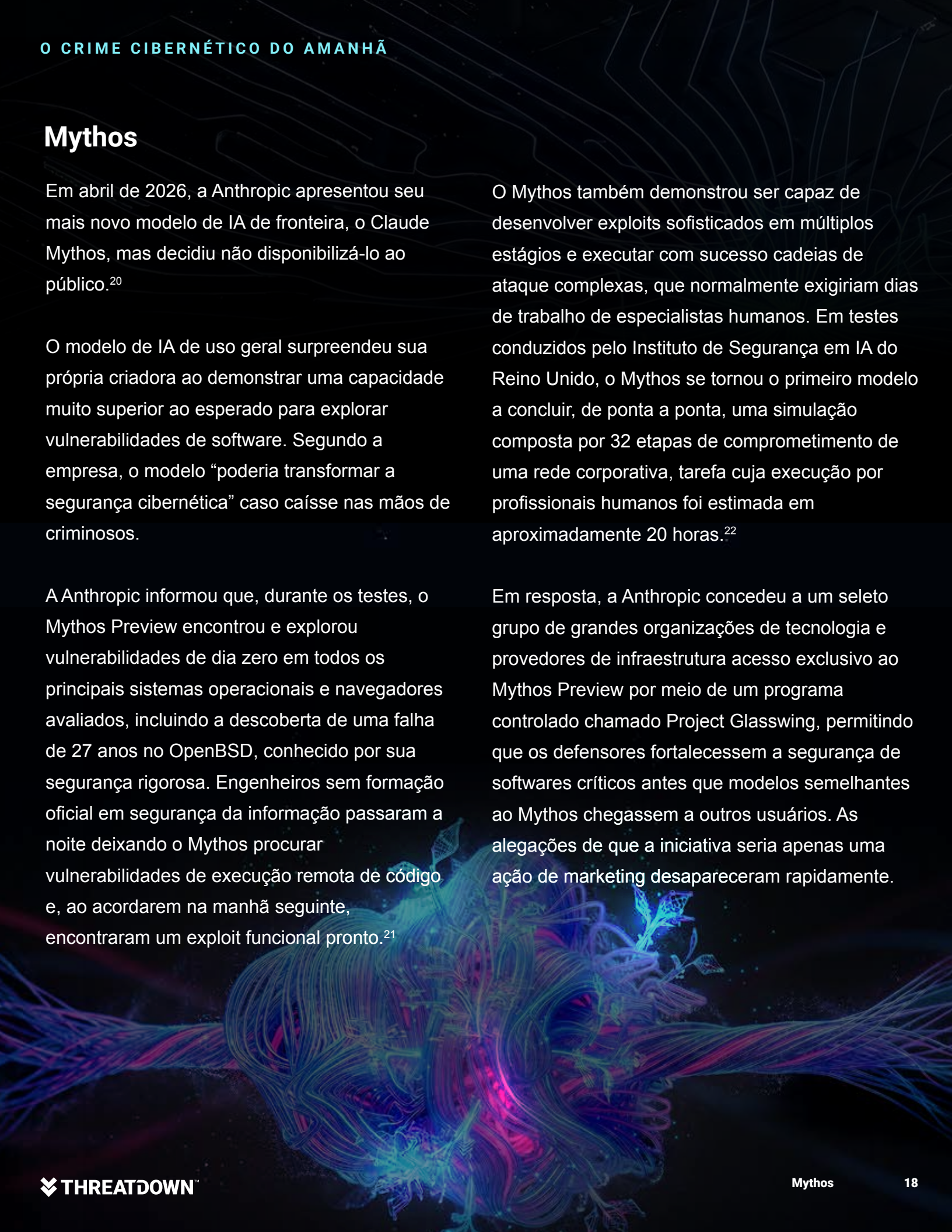
Em abril de 2026, a Anthropic apresentou seu mais novo modelo de IA de fronteira, o Claude Mythos, mas decidiu não disponibilizá-lo ao público.[20]

O modelo de IA de uso geral surpreendeu sua própria criadora ao demonstrar uma capacidade muito superior ao esperado para explorar vulnerabilidades de software. Segundo a empresa, o modelo "poderia transformar a segurança cibernética" caso caísse nas mãos de criminosos.

A Anthropic informou que, durante os testes, o Mythos Preview encontrou e explorou vulnerabilidades de dia zero em todos os principais sistemas operacionais e navegadores avaliados, incluindo a descoberta de uma falha de 27 anos no OpenBSD, conhecido por sua segurança rigorosa. Engenheiros sem formação oficial em segurança da informação passaram a noite deixando o Mythos procurar vulnerabilidades de execução remota de código e, ao acordarem na manhã seguinte, encontraram um exploit funcional pronto.[21]

O Mythos também demonstrou ser capaz de desenvolver exploits sofisticados em múltiplos estágios e executar com sucesso cadeias de ataque complexas, que normalmente exigiriam dias de trabalho de especialistas humanos. Em testes conduzidos pelo Instituto de Segurança em IA do Reino Unido, o Mythos se tornou o primeiro modelo a concluir, de ponta a ponta, uma simulação composta por 32 etapas de comprometimento de uma rede corporativa, tarefa cuja execução por profissionais humanos foi estimada em aproximadamente 20 horas.[22]

Em resposta, a Anthropic concedeu a um seleto grupo de grandes organizações de tecnologia e provedores de infraestrutura acesso exclusivo ao Mythos Preview por meio de um programa controlado chamado Project Glasswing, permitindo que os defensores fortalecessem a segurança de softwares críticos antes que modelos semelhantes ao Mythos chegassem a outros usuários. As alegações de que a iniciativa seria apenas uma ação de marketing desapareceram rapidamente.



Após aplicar uma versão inicial do Mythos à sua base de código, o Firefox lançou impressionantes 423 correções de bugs em abril de 2026, contra apenas 31 no mesmo mês do ano anterior.[23] O CTO da Mozilla descreveu a experiência como tendo provocado “vertigem” na equipe.[24]

Aparentemente em resposta ao Mythos, em 2 de junho, o governo dos Estados Unidos publicou uma ordem executiva determinando que a National Security Agency (NSA) desenvolvesse um processo confidencial de análise comparativa das capacidades de segurança cibernética de modelos de IA de fronteira. A medida criou uma nova classificação denominada “covered frontier model”, cujo critério seria definido por esse processo, e estabeleceu um mecanismo que permite ao governo acessar esses modelos por até 30 dias antes de sua disponibilização a outros parceiros de confiança. A ordem também criou um centro nacional de confiança. A ordem também criou um centro nacional de coordenação para segurança cibernética em IA, destinado a coordenar a verificação de vulnerabilidades e a distribuição de patches em infraestruturas críticas, além de determinar que a Cybersecurity and Infrastructure Security Agency (CISA) ampliasse os programas federais de fornecimento de ferramentas defensivas baseadas em IA.[25] Foi o sinal mais claro possível de que a segurança cibernética havia mudado para sempre.

Em 9 de junho, a Anthropic lançou o Fable 5, um modelo da classe Mythos “tornado seguro para uso geral”, e o Mythos 5, uma versão menos restritiva destinada exclusivamente a defensores cibernéticos e provedores de infraestrutura.[26] Três dias depois, o governo dos Estados Unidos publicou uma diretriz de controle de exportações suspendendo o acesso a ambos os modelos por qualquer cidadão estrangeiro, estivesse ele dentro ou fora do território nacional, após a descoberta de um jailbreak para o Fable 5. As restrições foram tão amplas que a Anthropic desativou os dois modelos para todos os clientes a fim de garantir conformidade com a medida.

> Segundo a empresa, o modelo **“poderia transformar a segurança cibernética”** caso caísse nas mãos de criminosos.

Entretanto, enquanto o Fable 5 e o Mythos 5 permaneciam indisponíveis, o desenvolvimento da IA continuou avançando. Em 22 de junho, a OpenAI ampliou sua resposta ao Project Glasswing por meio do programa Daybreak, com restrições semelhantes, e lançou o GPT-5.5-Cyber, um modelo de IA de fronteira otimizado para descobrir e explorar vulnerabilidades.[27] Outros certamente virão em seguida. O CEO da Anthropic estimou que modelos chineses com capacidades comparáveis às do Mythos provavelmente estarão amplamente disponíveis dentro de seis a doze meses.[28]

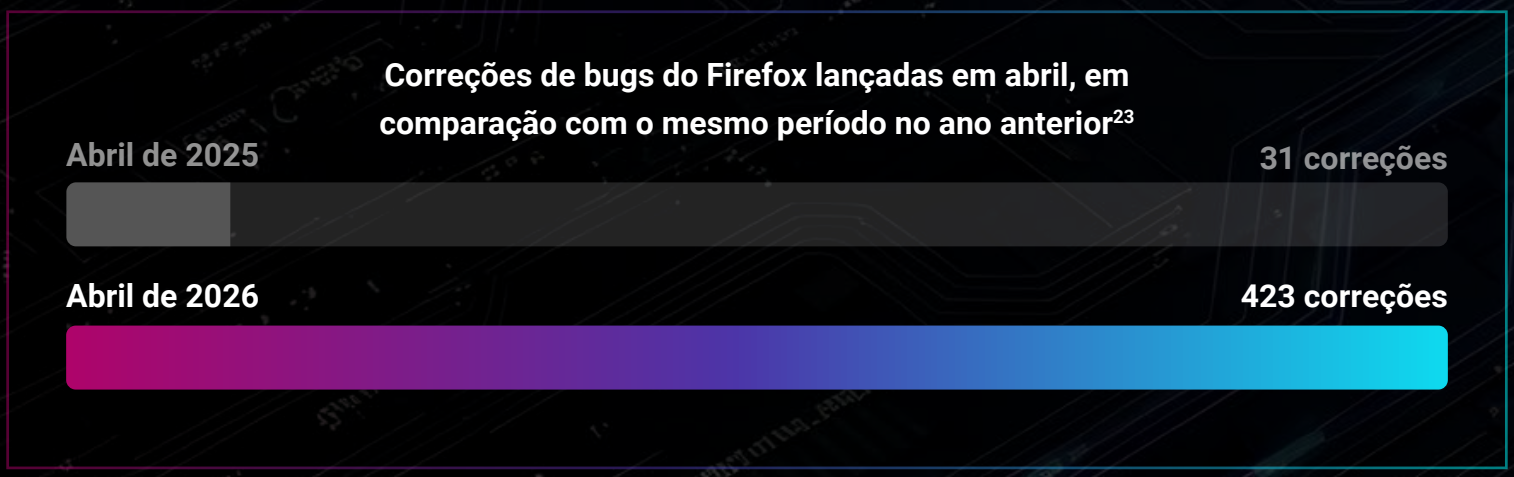
A suspensão permaneceu em vigor até 30 de junho, quando as restrições de exportação foram revogadas. O Fable 5 voltou a ficar amplamente disponível em 1º de julho. O Mythos 5 continua acessível apenas para defensores cibernéticos e provedores de infraestrutura previamente aprovados, conforme o projeto original da Anthropic, e não em decorrência de restrições governamentais.[29]

É apenas uma questão de tempo, talvez apenas de alguns meses, até que o ecossistema criminoso, que já comercializa acesso a modelos de IA de fronteira desbloqueados (jailbreak) e a modelos de código aberto sem mecanismos de proteção, passe também a oferecer modelos da classe Mythos.

Porém, no curto prazo, modelos dessa classe começam a chegar às mãos dos defensores, o que à primeira vista pode parecer uma boa notícia. Na prática, entretanto, a situação é mais complexa. Modelos capazes de descobrir vulnerabilidades nessa escala produzirão um volume de patches muito superior ao que as organizações conseguem administrar atualmente. Como cada novo patch representa uma oportunidade para que criminosos realizem engenharia reversa e desenvolvam um exploit, o número de exploits disponíveis para uso criminoso poderá aumentar significativamente a curto prazo.

Ao mesmo tempo, organizações que já enfrentam dificuldades para lidar com seu acúmulo atual de patches pendentes não passarão, de uma hora para outra, a ter dez vezes mais capacidade para aplicar dez vezes mais correções. Elas ficarão ainda mais atrasadas, e o período durante o qual os exploits permanecem eficazes tenderá a aumentar.

O volume de trabalho que modelos da classe Mythos criarão para os defensores é tão significativo quanto a ameaça que eles representam quando utilizados pelas pessoas erradas.

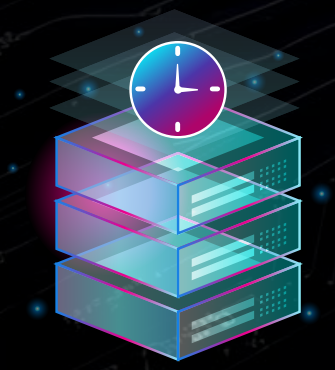


O volume de trabalho que modelos da classe Mythos criarão para os defensores é tão significativo quanto a ameaça que eles representam quando utilizados pelas pessoas erradas.

Conclusão

Tudo o que é apresentado neste relatório é concreto e real. Não há hipóteses nem cenários especulativos. O ataque ao serviço de abastecimento de água de Monterrey foi real. O mercado de IA criminosa, o primeiro exploit de uma vulnerabilidade de dia zero desenvolvido com IA, as dezenas de milhões de downloads de modelos de IA sem mecanismos de proteção e a skill comprometida no mercado do OpenClaw são todos fatos reais.

A questão que sua empresa enfrenta não é se os ataques nativos de IA representarão uma ameaça, mas se ela estará preparada quando esses ataques ocorrerem. A transição para um ecossistema de crime cibernético moldado pela IA já começou e talvez restem apenas seis meses para se preparar em três áreas centrais:



Realize correções como se o tempo estivesse acabando.

O Mythos e seus sucessores poderão inundar as organizações com um enorme volume adicional de atualizações de software. As organizações que conseguirem manter seus sistemas atualizados por meio de automação e processos eficientes de Vulnerability Assessment e Patch Management estarão em uma posição muito mais segura. As demais enfrentarão um acúmulo cada vez maior de sistemas sem correção, períodos mais longos de exposição e criminosos munidos de uma quantidade crescente de novos exploits.

Monitore continuamente, 24 horas por dia, 7 dias por semana.

A inteligência artificial torna cada invasor mais rápido, mais eficiente e mais escalável, mas não cria táticas novas. Os dados armazenados nos endpoints continuam sendo o principal alvo, e o roubo de identidades permanece sendo a principal forma de obtenção de acesso. As organizações precisam estar preparadas para reagir rapidamente aos primeiros sinais de um ataque impulsionado por IA, seja por meio de um Centro de Operações de Segurança (SOC), seja com um serviço de Managed Detection and Response (MDR).

Descubra e governe sua Shadow AI.

Ferramentas de IA não autorizadas, skills de agentes e conexões MCP representam um risco cada vez maior à conformidade, à privacidade e à segurança, e sua equipe de TI não consegue enxergá-los. Utilize soluções de detecção e resposta para IA para identificar a superfície d ataque da Shadow AI existente em sua organização.

O futuro do crime cibernético já chegou, só não está distribuído de forma uniforme ainda. As organizações que atravessarem esse momento sem sofrer grandes impactos serão aquelas que compreenderam essa transformação com clareza e agiram enquanto o tempo ainda favorecia quem estava preparado.

APÊNDICE

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