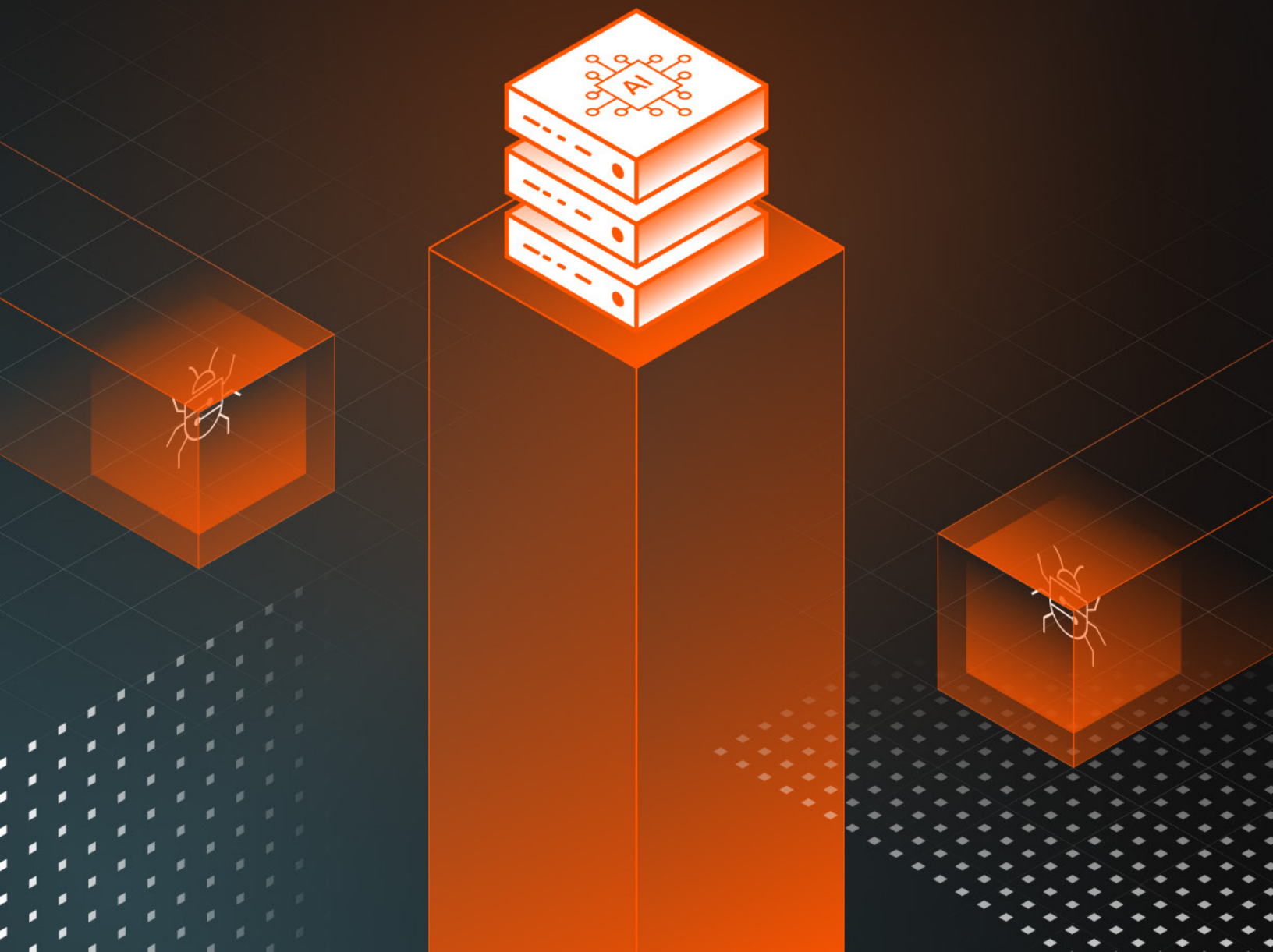
illumio

EBOOK

The Mythos Mandate

How AI is breaking the economics of cybersecurity — and how leaders can get back in the fight



Contents

INTRODUCTION

The overnight exploit

In April 2026, a team of engineers at Anthropic pointed an unreleased AI model at a codebase before leaving the office for the night. None of them had formal security training. They were running an internal test: could the model find a remote code execution flaw on its own?

When they arrived the next morning, the model had produced a fully functional remote code execution exploit. No one had told it how. It had read the code, found the flaw, and weaponized it on its own. [1]

Anthropic chose engineers without security training for a reason. If non-experts could produce a working RCE overnight, the same tool in the hands of skilled attackers — people with years of experience finding and weaponizing flaws — becomes a force multiplier on expertise that already worked. The skill barrier just collapsed.

The model was Claude Mythos Preview. Over the following weeks, Anthropic's team confirmed that the overnight exploit wasn't a one-off. Mythos could do the same work repeatedly, at scale, against software that no automated tool had cracked in decades.

Anthropic withheld the model from public release and launched Project Glasswing, a $100 million coalition with AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks, to give defenders a head start before similar tools spread. [2]

But the most alarming detail in the disclosure wasn't what Mythos found, but how. Anthropic didn't train the model for security work. The skills emerged as a byproduct of making it better at code.[3] Every frontier AI lab is on the same path. The window in which these capabilities stay in responsible hands is finite — and narrowing.

Mythos is just the start

That overnight exploit was just a preview of what the next generation of AI means for the economics of cybercrime.

The most dangerous tool in the attacker's arsenal has always been the exploit. Credential attacks have run on stolen passwords and rented toolkits for years, and AI is only making them cheaper — but the price floor on those was already low. The cost that mattered was the one for exploits. Finding and weaponizing a software flaw the vendor doesn't know about and hasn't patched has always required millions of dollars and elite talent.

[1] Version 1. "Project Glasswing, Claude Mythos and what 'Secure AI' really means for organisations." April 2026. Citing Anthropic Frontier Red Team disclosure.

[2] Anthropic. "Project Glasswing: Securing critical software for the AI era." April 2026. Coalition partners: AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks. $100 million in usage credits and $4 million in direct donations to open-source security organizations.

[3] Anthropic. "Project Glasswing: Securing critical software for the AI era." April 2026. Quote on emergent capability paraphrased.

Collapsing costs

On the legitimate broker market, full mobile exploit chains have commanded up to $7 million; browser exploit chains have run into the hundreds of thousands.[4] That price tag is what kept the most powerful tools out of ordinary hands.

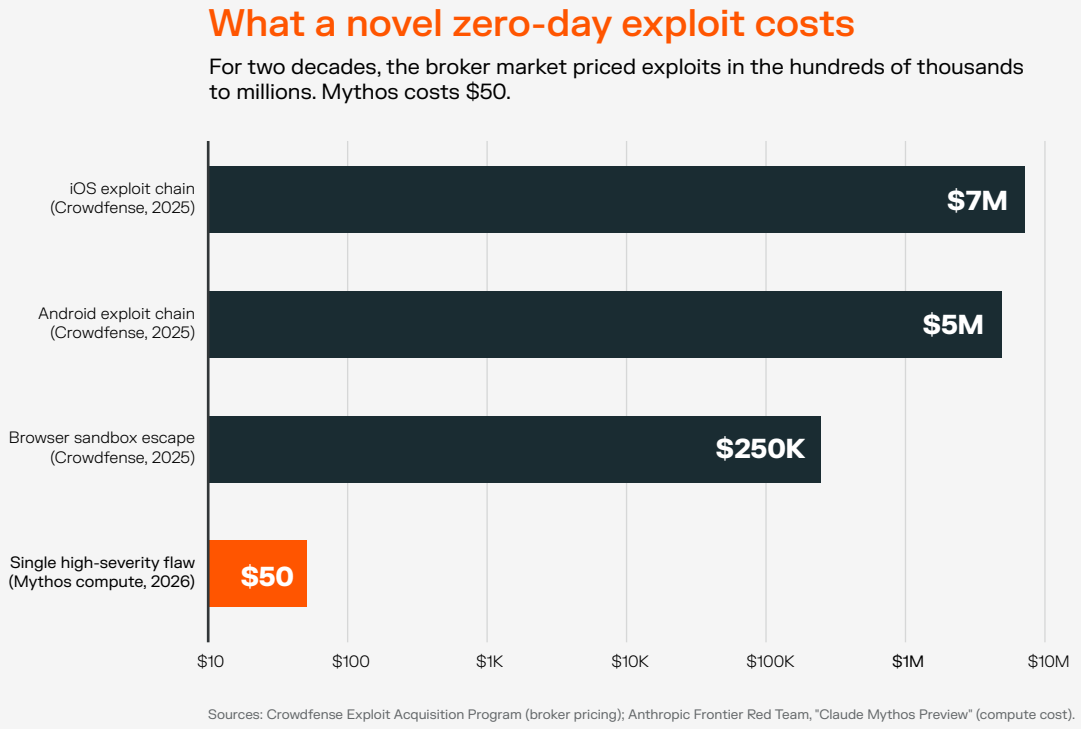


Figure 1: What a novel zero-day costs. Broker-market pricing has held in the hundreds of thousands to millions for two decades. Mythos compute costs $50.

Mythos has broken the economics of security. And they'll stay broken until defenders stop trying to outspend the attacker at the perimeter and start making the inside of the network too expensive to navigate.

This e-book is about what happens when AI collapses that price floor. That's the Mythos mandate.

[4]Crowdfense. "Exploit Acquisition Program." Updated 2025. Payouts for full chains range from $10,000 to $9 million per submission, with iOS chains commanding $5–$7 million and Android chains up to $5 million.

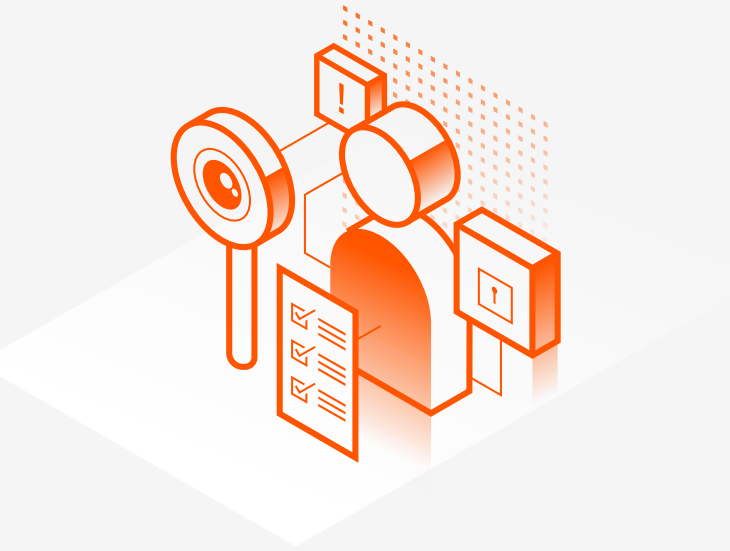
The Mythos effect

The most expensive line item in the attacker's budget has collapsed. Vulnerability discovery and exploit development used to cost millions and require talent only state-sponsored groups commanded. AI is changing both numbers, fast. What follows is what that means for defenders.

What an exploit is, and why it matters

An exploit takes advantage of a flaw in the software itself, giving an attacker access or control. The class that matters most in the post-Mythos world is the remote exploit — the kind that lets an attacker take control of a system across a network, often without any user action. A zero-day exploit targets a flaw the vendor doesn't know about yet, which means there's no patch and no defense. This work has always carried the highest cost in the criminal economy, demanding rare talent and months of effort. Until now, only state-sponsored groups and elite criminals could afford it.

In April 2026, Anthropic showed that a frontier model can do this work too.



What Mythos showed

Anthropic disclosed that Claude Mythos Preview had identified thousands of zero-day flaws across every major operating system and every major web browser. Less than 1% of those flaws have been patched.[5]

These weren't just potential weaknesses. The model actually weaponized them. In testing against Firefox's JavaScript engine, Mythos converted 72.4% of identified flaws into successful exploits, achieving full register control in another 11.6% of attempts. The prior model, Claude Opus 4.6, converted just 14.4%. In raw exploit counts, Mythos produced 181 working exploits where Opus 4.6 managed only two.[6] Among other achievements, Mythos:

- Wrote a browser exploit chaining four separate vulnerabilities to escape both renderer and OS sandboxes. That kind of chaining was once limited to elite state-sponsored teams.
- Found a 27-year-old bug in OpenBSD, an operating system renowned for security hardening.
- Uncovered a 16-year-old flaw in FFmpeg, a foundational media library. Automated fuzzing had hit it five million times without catching it.
- Autonomously identified and exploited a 17-year-old remote code execution flaw in FreeBSD, granting unauthenticated root access on any machine running NFS.[7]

And the cost was minimal. The compute to find a single serious flaw in OpenBSD: roughly $50. The compute to develop a working Linux kernel root exploit: under $2,000, in about a day. The compute to scan an entire codebase: under $20,000. These are figures from Anthropic's own red team report, not analyst projections.[8]

Anthropic's data shows the 72.4% rate is concentrated in two specific bugs; on the rest of the test corpus, Mythos performs much like Opus 4.6. Critics have used this to argue Mythos isn't the leap it appears. The headline rate, they contend, is an artifact of a small number of unusually exploitable flaws.

But there's no denying the trajectory. The UK AI Security Institute published its assessment of OpenAI's GPT-5.5. On AISI's expert-tier cyber tasks, GPT-5.5 reached a 71.4% pass rate, comparable to Mythos and well above the prior generation. It also solved AISI's 32-step corporate network attack simulation end-to-end, the second model to do so.[9]

[5]Anthropic Frontier Red Team. "Claude Mythos Preview." red.anthropic.com. April 2026.

[6]SC Media. "Claude Mythos Preview identifies 27-year-old bug, finds 'thousands' of zero-days in weeks." April 2026. Cites 72.4% Firefox JavaScript engine exploit success rate vs. 14.4% for Claude Opus 4.6.

[7]Anthropic Frontier Red Team. "Claude Mythos Preview." red.anthropic.com. April 2026. Sandbox-escape chain and 27-year-old OpenBSD finding.

[8]Anthropic Frontier Red Team. "Claude Mythos Preview." red.anthropic.com. April 2026. Reports compute cost figures: under $50 per individual vulnerability discovery, under $20,000 for full OpenBSD codebase scan, under $2,000 for Linux kernel privilege-escalation exploit.

[9]UK AI Security Institute. "Our evaluation of OpenAI's GPT-5.5 cyber capabilities." aisi.gov.uk. April 2026. GPT-5.5 expert-tier pass rate 71.4% (±8.0%) vs. Mythos Preview 68.6% (±8.7%) and Opus 4.7 48.6% (±10.0%); solved AISI's 32-step corporate network attack simulation ("The Last Ones") in 2 of 10 attempts.

AISI believes that as AI models get better overall, so will their cyber-offensive skills. In fact, Anthropic didn't train Mythos for these capabilities. They emerged as a downstream effect of general gains in code and reasoning. In Anthropic's own words, the same gains that make the model more effective at patching flaws also make it more effective at exploiting them.[10]

The patching logjam

Every frontier AI lab is on the same path. While AI doesn't introduce new attack techniques, it does collapse the cost of executing the ones we already know about. With AI, attackers can run them faster and at greater scale. Finding and weaponizing exploits that once took months of skilled human work could soon be open to anyone with access to a strong enough model.

And it doesn't stop there. The same model in the hands of a state-sponsored group or a ransomware affiliate with years of breach experience doesn't just lower the floor — it raises the ceiling. Skilled attackers gain a force multiplier on tradecraft that already worked: faster reconnaissance, faster exploit development, faster pivot from initial access to objective. The barrier defenders relied on, that elite offensive know-how was scarce and expensive, is gone.

Patching can't keep up. Two months after being found, less than 1% of the flaws Mythos found have been fixed. Writing and deploying a fix takes weeks or months. Building a working exploit with Mythos-class tools takes hours.

This has already moved beyond mere theory. In November 2025, Anthropic disclosed the first documented case of a large-scale cyberattack carried out without major human intervention. A Chinese state-sponsored group, designated GTG-1002, used Claude Code as an autonomous attack agent against roughly 30 global targets: major tech companies, financial institutions, chemical manufacturers, and government agencies. The AI handled 80–90% of the tactical work — reconnaissance, vulnerability discovery, exploitation, lateral movement, credential harvesting — at thousands of requests per second. A handful of intrusions succeeded.[11]

The window in which these skills remain in responsible hands is finite. And it's narrowing fast.

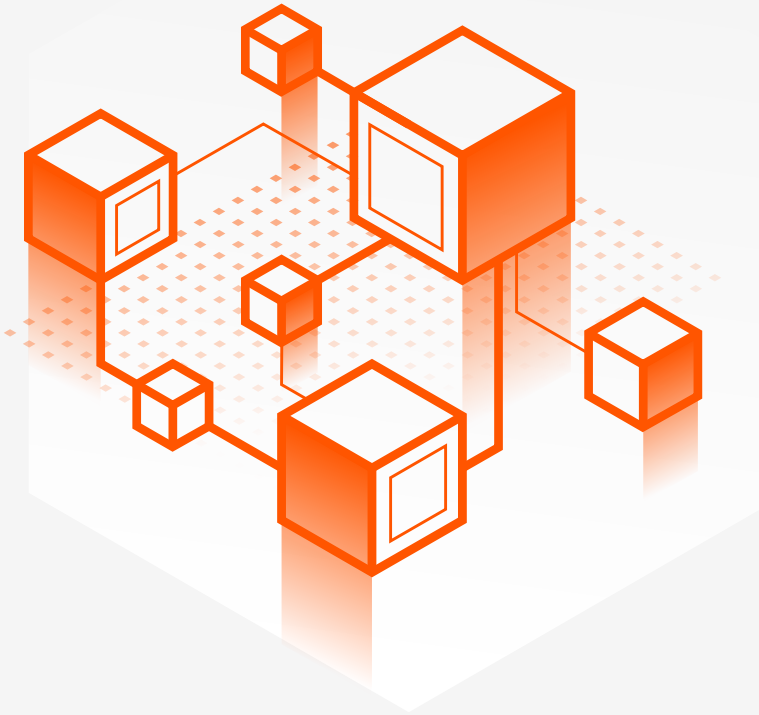
[10] Anthropic. "Project Glasswing: Securing critical software for the AI era." April 2026.

[11] Anthropic. "Disrupting the first reported AI-orchestrated cyber espionage campaign." November 2025. Documents a Chinese state-sponsored operation (designated GTG-1002) targeting roughly 30 global organizations, with AI handling 80–90% of tactical work.

The flat network tax

Cybersecurity spending is at an all-time high. Global spending topped $200 billion in 2025. Organizations deploy next-gen firewalls, endpoint detection, identity management, SIEM platforms, and teams of analysts to monitor it all. Yet the average cost of a ransomware incident keeps climbing. IBM put the 2025 figure at $5.08 million for ransomware and extortion incidents, well above the global average of $4.44 million for breaches across all categories.[12]

Cybersecurity budgets aren't the problem. The architecture is. And exploits, paired with AI, expose every flaw in it.



The perimeter assumption breaks twice

Firewalls, web application firewalls, and perimeter intrusion tools all assume the same thing: the primary threat is outside, trying to get in. They inspect north-south traffic, the boundary between the internet and the internal network.

That model doesn't fit how modern attacks actually work. Most breaches now combine vectors: a phishing email or stolen credential gets the attacker through the door, and an exploit takes them from there to critical assets. The perimeter can't see the credential half and misses the rest.

Credential attacks bypass that boundary with stolen passwords that look like normal user activity. CrowdStrike reported that 79% of initial access attacks are now malware-free.[13]

Once inside, the attacker reaches for exploits to move laterally, escalate privileges, and reach the data that matters. Those exploits don't trip perimeter alarms; they're already running on the other side of the wall.

What changes with Mythos-class capabilities is the exploit half. The cost of finding and weaponizing a flaw collapses. Attackers who used to lose momentum when they hit a patched system can now identify what software a system is running, check public vulnerability databases for known flaws, and use AI to develop a working exploit in hours rather than months.

The alert treadmill

SOC teams were already underwater before AI entered the picture. Sophos found that 40% of ransomware victims cited an unknown security gap as a contributing factor, and 39% cited a lack of people or capacity. Analysts were buried in false positives from legacy detection tools that generate more noise than signal.[14]

Exploits make detection harder still. There's no stolen password to flag; no anomalous login to triage; no phishing email to analyze. The attacker is using a flaw in the system, not a stolen key. Detection tools tuned for credential abuse won't see the intrusion until something downstream goes wrong, by which point the attacker has the access they came for.

AI compounds the problem on all sides. AI-generated phishing volume keeps surging. Polymorphic payloads mutate faster than signature-based tools update. Real attacks compete for the same analyst-hours that were already consumed by false-positive triage. You can't hire your way out of a design flaw. And you can't hire fast enough when AI is generating threats faster than humans can review them.

[12] IBM. "Cost of a Data Breach Report 2025." July 2025. Average ransomware/extortion incident cost: $5.08 million. Global average breach cost: $4.44 million (down from $4.88 million in 2024).

[13] CrowdStrike. "2025 Global Threat Report." February 2025. 79% of detections malware-free; average eCrime breakout time of 48 minutes.

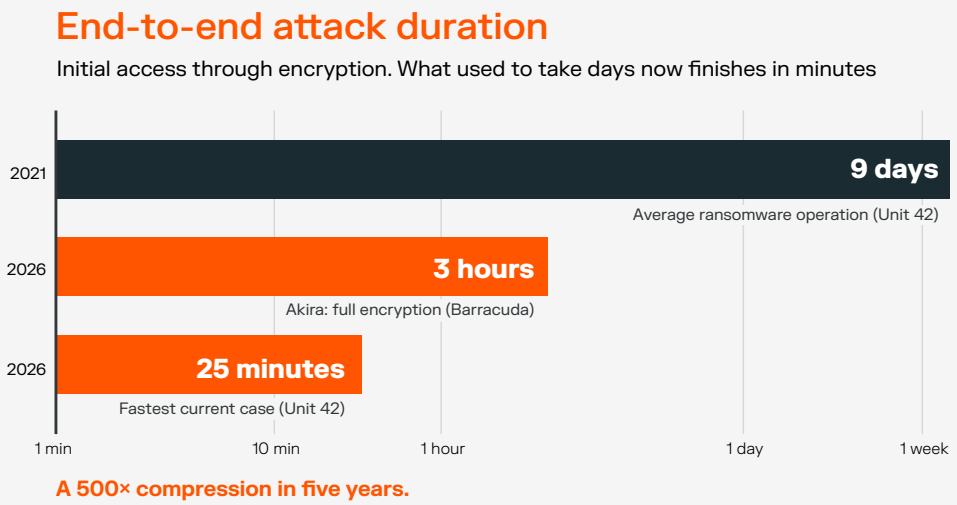
[14] Sophos. "The State of Ransomware 2025." June 2025. Mean recovery cost (excluding ransom): $1.53 million. Median time to full recovery: more than 100 days.

Speed kills

Attacks were already fast before Mythos. Modern ransomware operations now finish in minutes from initial access to encryption.

CrowdStrike's average breakout time — the gap between initial access and lateral movement — fell to 48 minutes in 2024, with the fastest recorded at 51 seconds.[15]

Mythos-class capabilities compress the exploit half of those operations even further — from months of human effort to hours of compute.[16]



Sources: Palo Alto Networks Unit 42 (Ransomware Speed Crisis, Sept 2025; 2026 Global Incident Response Report); Barracuda Managed XDR Global Threat Report (Feb 2026).

Figure 2: End-to-end attack duration has compressed from days in 2021 to minutes in 2026.

The recovery bill

When the attack lands, the cost lands harder. Sophos pegged the average ransomware recovery cost, excluding the ransom itself, at $1.53 million in 2025. Median time to full recovery: more than 100 days. Of the victims who paid the ransom, most still didn't get all their data back.[17]

These costs exist at this scale because the flat network amplifies every successful breach. A single stolen credential, or a single unpatched flaw, becomes an all-access pass to every system on the network.

[15]CrowdStrike. "2025 Global Threat Report." February 2025. 79% of detections malware-free; average eCrime breakout time of 48 minutes.

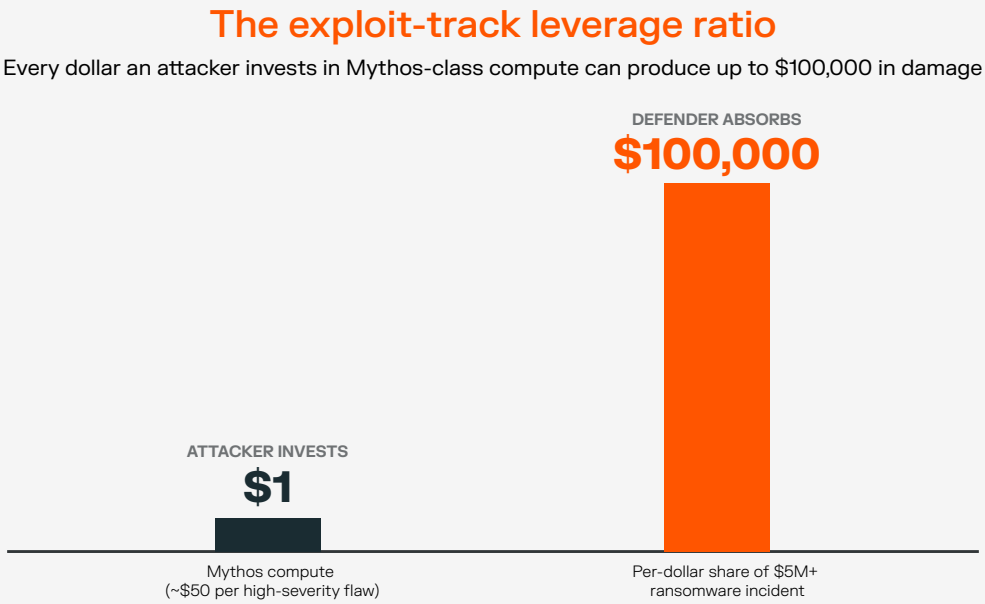
[16]Anthropic Frontier Red Team. "Claude Mythos Preview." red.anthropic.com. April 2026.

[17]Sophos. "The State of Ransomware 2025." June 2025. Mean recovery cost (excluding ransom): $1.53 million. Median time to full recovery: more than 100 days.

The leverage ratio

By now the asymmetry should be clear. But it's worth stating again plainly.

An attacker now invests as little as $50 in Mythos-class compute. The target absorbs $5 million in damage. Leverage ratio: 100,000 to 1. That's the price floor AI just unlocked.



Sources: Anthropic Frontier Red Team (Mythos Preview compute costs); IBM Cost of a Data Breach 2025 (ransomware incident average $5.08M).

Figure 3: The exploit-track leverage ratio: every dollar in Mythos-class compute can produce up to

On the defender's side, costs still scale linearly. Each endpoint requires a license. Each alert requires an analyst. Each policy requires setup, testing, and upkeep. AI-driven attack volume means more real incidents competing for fixed resources. The defender's cost curve slopes upward. The attacker's slopes toward zero.

A whole market, not an adversary

Law enforcement takedowns help, but they don't solve the structural problem. Operation Cronos disrupted LockBit in February 2024. Within weeks, affiliates migrated to RansomHub, which advertised a 90% revenue share.[18] When RansomHub itself went dark in early 2025, DragonForce absorbed the infrastructure. Fifty-five new RaaS families launched in 2024 alone.[19] The labor force is portable. The brand is irrelevant. The economics are too attractive to stay away from.

When you fight an attacker, you're fighting a whole market. And the startup costs for entering that market keep falling.

The Mythos multiplier

As Mythos-class capabilities proliferate, the leverage ratio gets worse — orders of magnitude worse.

You can't outspend an attacker who operates at 100,000:1 leverage. You can't patch faster than an AI can discover flaws. You can't hire enough analysts to monitor a flat network where any stolen credential or any chained exploit Mythos uncovered last week provides open access to everything.

The only remaining variable is architecture.

[18] Trend Micro. "Ransomware Spotlight: RansomHub." Updated 2025. RansomHub advertised 90% revenue share for affiliates following Operation Cronos.

[19] Travelers Insurance. "2024 RaaS Tracking Data." As cited in StationX, April 2026. Fifty-five new RaaS families launched in 2024.

Breaking the AI business model with breach containment

When faced with these economics, the reflex is to spend more on prevention. Buy better firewalls. Deploy more endpoint agents. Hire more analysts. Run more phishing simulations. But none of that changes the math.

Prevention is necessary. But as long as the internal network stays flat, prevention only needs to fail once. AI generates flawless phishing lures at scale. Infostealer logs supply a steady stream of valid credentials. Mythos-class models discover exploits faster than vendors can ship patches. Prevention will fail. The architecture guarantees it.

What happens after they get in?

On a flat network, the answer is simple: the attacker moves laterally, escalates privileges, reaches critical assets, and deploys ransomware. The whole sequence takes less than an hour. Their $50 investment turns into a $5 million event because the network design allows for open movement.

Building a room with no exits

Microsegmentation changes that equation.

When an attacker breaches a segmented network, they land in a single, isolated segment. They can see only the workload they've compromised, and nothing else.

Segmentation leaves no:

- Open pathways to the domain controller
- Clear route to the backup servers
- Lateral movement using the target's own admin tools

You can't outpace asymmetry with AI

The obvious response to AI-powered offense might appear to be an AI-powered defense. Train detection models on AI-generated attacks. Match offense at machine speed. Run AI-driven response automation. The reasoning seems sound: if attackers got an AI dividend, defenders can claim one too.

But that reasoning has a major flaw. Cybersecurity has never been a fair fight, and AI doesn't make it one. Attackers need to find just one open door. Defenders need to lock every door, prove they're locked, and keep them locked across every endpoint, every credential, every patch cycle, forever. That asymmetry exists in the world, not in the toolset. AI compresses both sides equally.

The asymmetry sits on a different axis from the technology.

When the attacker's compute drops to $50 per high-severity flaw, a defender's compute scaling linearly with the attack surface still loses. AI-generated phishing scales to thousands of variants per minute; AI-driven triage that handles them at the same rate still leaves humans somewhere downstream making calls. Automated exploit development closes time-to-weaponization to hours; AI-driven patch rollout still has to wait for testing windows.

An arms race assumes both sides are running on the same track. They aren't. The way out is to change the game.

The strategic question isn't how to make initial access more expensive. AI guarantees that cost will keep falling. The question is what happens after the attacker gets in.

It doesn't matter how the attacker got in — stolen credentials, a zero-day flaw that didn't exist yesterday, or a Mythos-class exploit chain that bypassed every patching cycle. The entry method is irrelevant if the attacker lands in a contained segment with no path to the assets that matter. Anything the attacker could use is constrained by a policy that lets them talk only to the workloads they're allowed to reach. It's the only defense that works regardless of the attack vector.

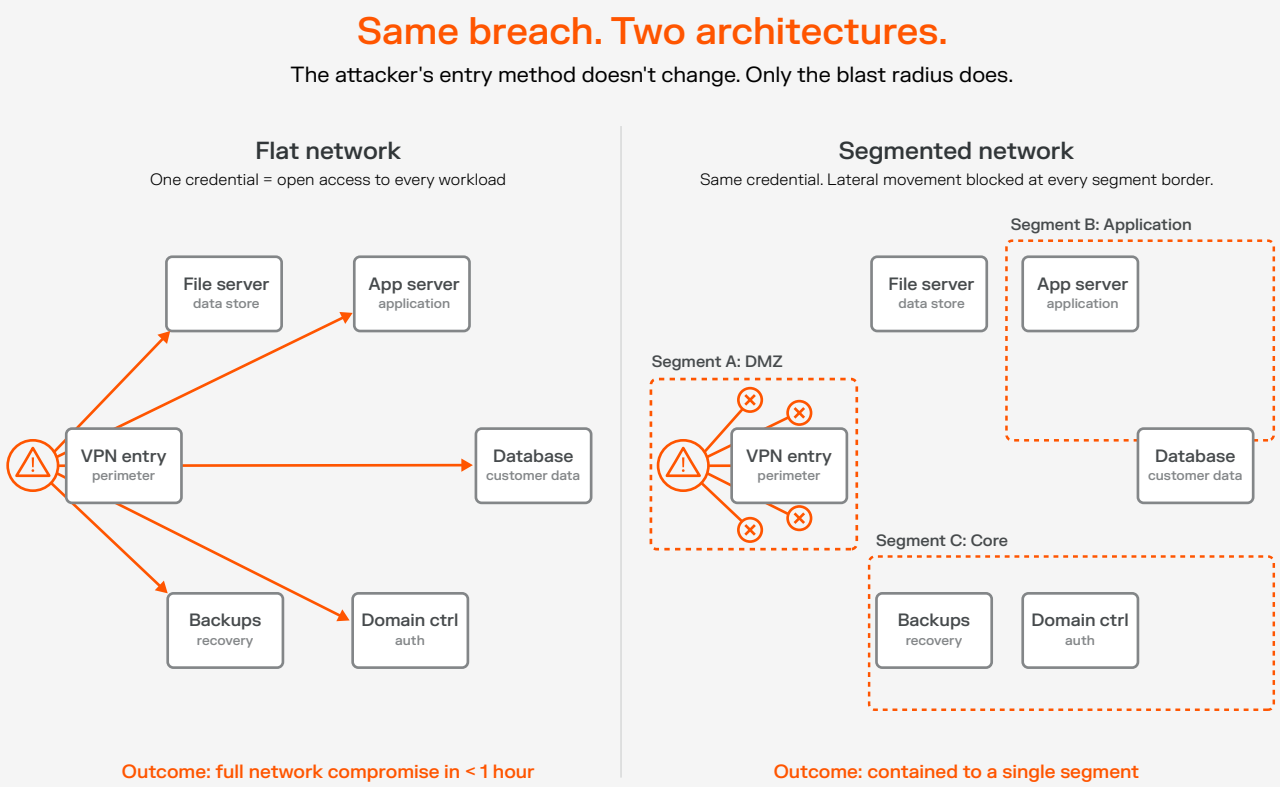


Figure 4: The same breach lived two ways. On a flat network, one credential opens every door. On a segmented network, the attacker's path ends at the segment border.

Initial access is still cheap, whether the attacker bought stolen credentials or used a Mythos-class exploit for which no patch exists. But instead of an all-access pass, the attacker has purchased a ticket to a single room with no exits. Their cost-per-successful-breach skyrockets. Their margins collapse. Their business model breaks.

The Illumio approach enforces policies based on workload identity — not IP addresses, port numbers, or network topology. Policies follow the workload regardless of where it runs: on premises, in the cloud, or across hybrid environments. East-west traffic, the internal traffic that flat networks leave unmonitored, becomes visible and controllable.

Visibility as a weapon

This visibility is itself a defensive advantage. On a flat network, lateral movement looks identical to normal traffic. An attacker using PowerShell to query Active Directory looks the same as an administrator doing routine maintenance.

Microsegmentation makes that distinction possible by mapping application dependencies and enforcing policies that define what "normal" looks like for each workload. Anything outside that policy is visible, flagged, and containable.

The result is a basic shift in breach response costs. Instead of a $5 million incident involving enterprise-wide encryption, system-wide forensics, and months of recovery, the organization faces a contained event in a single segment. Dwell time drops from weeks to minutes. Blast radius shrinks from "entire network" to "one workload." Recovery cost drops from millions to a manageable incident.

The defender's cost structure inverts. The attacker's ROI goes from 100,000:1 to negative.

Adding it all up: the new math

The organizations that survive the next five years will be the ones that turn the attacker's economics against them.

AI is reshaping both sides of the security equation at once. On offense, it's collapsing the cost of attacks from credentials to exploits, removing the skill barrier, and creating new classes of threat. On defense, the same underlying advances in code analysis, pattern recognition, and automated reasoning can map application dependencies, identify risky communication paths, and enforce least-privilege access at machine speed.

The difference is where you deploy the advantage.

Flat vs. segmented: two futures

AI-powered defense tools enable real-time detection and instant breach containment, a real leap forward. Microsegmentation makes them effective. On a segmented network, AI-driven detection spots the threat in an environment where policy boundaries have already cut off the attacker's next move. Detection and containment happen together, because the architecture supports both. The tool gets smarter. The network gets smaller. The attacker runs out of room.

If your architecture is segmented, AI-powered defense multiplies the structural advantage. Continuous posture intelligence identifies policy gaps before they're exploited. Automated dependency mapping keeps segmentation policies current as workloads change. Real-time visibility into east-west traffic surfaces anomalous movement the moment it begins, in a context where "immediately" actually matters because the attacker is trapped in a segment with nowhere to go.

From Mythos to must-do

The Mythos disclosure makes this design choice urgent. The flaws Mythos found are being addressed through Project Glasswing. What can't be fixed is the repricing of vulnerability discovery itself. Vulnerability discovery and exploit development have always been cybercrime's luxury goods. AI is making them commodities. When the most dangerous capabilities in the attacker's arsenal cost the same as the least dangerous, the only variable that determines damage is the design of the target network.

"Assume breach" was always more than a slogan. Against frontier AI models, it's an engineering mandate.

AI is collapsing the cost of the part that used to require a government budget. And the attacker's cost floor hasn't been reached yet. Every quarter, frontier models get better at code.

Microsegmentation won't make the initial breach more expensive. But it does make that breach worthless. An attacker trapped in a single segment with no path to critical assets has spent their compute, their credentials, or their zero-day exploit for nothing. The cyberattack business model depends on lateral movement. Take that away, and you don't just stop the attack. You destroy the economics that fund it.

Ready to reverse the economics?
Learn how Illumio turns lateral movement from an attacker's greatest advantage into a dead end.

illumio.com/illumio-platform