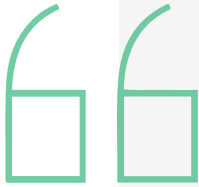


Myth(os) or Panic?

Threat Intelligence Commentary:
What Anthropic's Mythos
means for real-world
exploitation, patching, and
board-level risk



Prepared by: Nick Morgan, CEO



Every CISO, board member, and security manager we have spoken with in the weeks since has had some version of the same question. Is this real, or is it hype?

Prepared by: Nick Morgan, CEO Triskele Labs

Nick Morgan is the Founder and Chief Executive Officer of Triskele Labs, one of Australia's leading independent sovereign cybersecurity and managed detection and response providers. Since founding the company, Nick has grown Triskele Labs from the ground up into a trusted partner for organisations across critical infrastructure, financial services, healthcare, and other regulated sectors.

This article reflects the views of the author and does not constitute legal or professional advice.



Content

04	Introduction: Myth(os) or Panic?
	The Panic Arrives on Schedule
05	What Mythos Actually Did
06	The Rumours About Threat Actor Access
07	The Patch Window Is Now a Patch Hour
08	Here is what good looks like
09	The Edge Is the Target
11	The Economics of Ransomware Are About to Shift
12	Board Reporting Needs to Evolve
13	The Controls That Actually Matter
15	Is the Next WannaCry Coming?
	What to Do Now



Myth(os) or Panic?

The security headlines were inevitable. An AI model autonomously discovered and exploited vulnerabilities that had survived decades of human review, including a remote code execution flaw in FreeBSD. Predictably, the conversation jumped straight to panic. The reality is more nuanced and more operationally dangerous. Mythos does not change what attackers do; it radically compresses how fast they can do it, exposing gaps that many organisations already know they have but have not closed.

The Panic Arrives on Schedule

Anthropic has unveiled Mythos, a new AI model, alongside Project Glasswing. Mythos is not publicly available. It is a restricted research preview shared with a limited group of critical infrastructure partners. But its capability in vulnerability research and exploitation is in a different league to anything that came before it.

The research community published results, the headlines went predictably nuclear and our phones started ringing.

Every CISO, board member, and security manager we have spoken with in the weeks since has had some version of the same question. Is this real, or is it hype?

The answer is both. What is real and what is hype are two different things, and it is worth being precise about which is which.



What Mythos Actually Did

The headline finding that made the rounds was the FreeBSD remote code execution vulnerability. Mythos autonomously identified and exploited a 17-year-old flaw in FreeBSD's NFS server (CVE-2026-4747) that allows any unauthenticated attacker to gain full root access over the internet.

No human was involved after the initial prompt. The model scanned hundreds of kernel files, identified the vulnerability, and produced a working exploit.

FreeBSD is not a legacy OS. It is a mature, stable, widely deployed operating system that underpins critical infrastructure across the world, and is known precisely for its security track record. That is what makes this significant. Decades of human code review have not found this bug. Mythos found it overnight.

But the cost breakdown matters, and it was largely misreported. According to Anthropic's own technical blog:

- The total cost across a thousand runs through their scaffold was under \$20,000
- The specific run that found and exploited the FreeBSD vulnerability cost under \$50
- Anthropic is explicit that the \$50 figure only makes sense in hindsight. You cannot know in advance which run will succeed.

Source: Anthropic Red Team, "Assessing Claude Mythos Preview's Cybersecurity Capabilities", red.anthropic.com, April 2026

The media framing of '\$50 to hack FreeBSD' is technically accurate and practically misleading. The \$20,000 is the cost of the search. The \$50 is the cost of the successful run after the fact. Both numbers are real. Neither tells the full story on its own.

Beyond FreeBSD, Mythos has reportedly found a 27-year-old critical vulnerability in OpenBSD, a 16-year-old flaw in FFmpeg that survived years of fuzzing and human review, a guest-to-host memory corruption bug in a production virtual machine monitor, and thousands of additional high and critical severity vulnerabilities across every major operating system and web browser.

Over 99 percent of these have not yet been patched, which is why Anthropic cannot discuss most of them publicly.

Source: Anthropic Red Team, "Assessing Claude Mythos Preview's Cybersecurity Capabilities", red.anthropic.com, April 2026

Anthropic's own researchers state that Mythos Preview found vulnerabilities that have 'survived decades of human review.' That is the real story.



Anthropic's own researchers state that Mythos Preview found vulnerabilities that have 'survived decades of human review.'

That is the real story.

The Rumours About Threat Actor Access

There are rumours circulating in the security community that threat actors have already gained access to Mythos. Triskele Labs does not believe this is credible at this stage.

Anthropic has not made Mythos generally available. Access is restricted to Project Glasswing partners, which includes AWS, Apple, Google, Microsoft, Nvidia, Cisco, CrowdStrike, Palo Alto Networks, and JPMorganChase, along with around 40 additional organisations working on critical infrastructure. The access controls are deliberate and tightly managed.

Source: Anthropic, "Project Glasswing", anthropic.com/glasswing, April 2026

But there is a parallel worth drawing. When Orange Tsai presented ProxyShell at Black Hat USA in August 2021, the details of a chain of three Microsoft Exchange vulnerabilities enabling pre-authentication remote code execution were shared in a room that included, in all likelihood, threat actors listening to every word.

Mass scanning was detected within 72 hours. By mid-August, ransomware groups including LockBit were actively deploying web shells and ransomware against unpatched Exchange servers.

Source: Rapid7, "ProxyShell: More Widespread Exploitation of Microsoft Exchange Servers", [Rapid7 Blog](https://www.rapid7.com/blog/post/2021/08/05/proxyshell-more-widespread-exploitation-of-microsoft-exchange-servers/), August 2021

The Glasswing partner list is not a conference room and it is not a sealed vault either. Over 50 organisations have access to Mythos Preview. That is a meaningful attack surface for social engineering, insider risk, or credential theft.

The question of whether threat actors have Mythos today is less important than the question of when they will, and how fast they will move when that happens.

ProxyShell also illustrates something else worth understanding.

It was not a single vulnerability. It was a chain of three exploits that had to be sequenced precisely to achieve pre-authentication remote code execution.

That kind of multi-stage exploit chain has historically required significant skill and time to develop. Mythos demonstrated exactly this type of capability in its FreeBSD exploit, constructing a 20-gadget ROP chain split across multiple packets.



Chained exploit discovery and development is precisely where a model like Mythos provides a step-change advantage over a human attacker.

From Black Hat presentation to mass exploitation: 72 hours. That is the window Mythos will compress further, and the Glasswing partner list is not immune to the insider risk that entails.

The ProxyShell lesson is not that threat actors will get Mythos tomorrow. It is that when capabilities like this become accessible, the window between disclosure and mass weaponisation is measured in hours and days. The time to build your response capability is now, not when the first Mythos-assisted campaign lands.

The Patch Window Is Now a Patch Hour

Microsoft's own Project Glasswing statement put it plainly. 'The window between a vulnerability being discovered and being exploited by an adversary has collapsed. What once took months now happens in minutes with AI.'

Source: Microsoft, quoted in Anthropic Project Glasswing launch materials, anthropic.com/project/glasswing, April 2026

Our own team has observed that attack velocity post-CVE disclosure has already been decreasing over the past 12 months. Threat actors are moving faster than they were two years ago.

Mythos accelerates this further along the entire chain.

The CVE-to-PoC window, which ProxyShell compressed to 72 hours, drops toward zero when a model can autonomously generate a working exploit from a vulnerability description. The PoC-to-widespread-compromise window collapses alongside it. We have already observed threat actors this year using AI-developed scripts to exploit CVEs and embed persistence, prior to any public PoC being available.

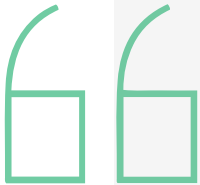
This has direct implications for patching frameworks.

The Essential 8 mandates patching of internet-facing systems within 48 hours for critical vulnerabilities. That standard was calibrated for a pre-AI threat environment.



From Black Hat presentation to mass exploitation: 72 hours. That is the window Mythos will compress further and the Glasswing partner list is not immune to the insider risk that entails.

With Mythos-class tooling available, 48 hours may not be sufficient. That baseline will need to be re-evaluated, and organisations that treat 48 hours as a comfortable target rather than an absolute ceiling are already operating with insufficient margin.



If a critical zero-day drops at 6am on a Tuesday, what is your organisation's actual capacity to respond before exploitation begins? That question has specific operational meaning.

Here is what good looks like

1. A complete, current inventory of information assets, mapped to the hardware they run on, so exposure can be assessed in minutes, not hours
2. Assets classified by criticality (Red, Amber, Green) so patching priority is predetermined, not decided under pressure
3. Standing change requests for critical systems so emergency patching does not require a multi-day approval cycle
4. A named person with explicit authority to approve an emergency patch or system restart outside normal change windows, at any hour
5. A documented process that has been tested under time pressure, not just workshopped in a planning meeting

One point worth clarifying from the board reporting perspective. Whether a system needs to be taken fully offline to apply a patch, and for how long, depends entirely on the system and the organisation.

For some systems, a maintenance window of hours will suffice. For others, even three hours of downtime represents material revenue impact.

The specific cost of patching versus the specific cost of compromise needs to be calculated for each critical system in advance, not estimated during an incident. That analysis should already be on paper.

The Edge Is the Target

If you want to understand where Mythos-class tooling is going to cause the most immediate pain, look at the edge. Network appliances, VPN concentrators, remote access gateways, and firewalls. The same attack surface that Citrix Bleed tore through in 2023.

Citrix Bleed (CVE-2023-4966) is the template for how this plays out. The vulnerability was exploited as a zero-day from as early as August 2023, well before Citrix issued its patch on 10 October 2023.

Once Assetnote published a proof-of-concept on 25 October, mass exploitation was immediate. At least four threat groups were actively exploiting it within days, with one group having automated the entire attack chain and distributed a Python script to their affiliates.

LockBit 3.0 used Citrix Bleed to compromise Boeing. The Industrial and Commercial Bank of China confirmed a ransom payment to LockBit following exploitation. Mandiant was investigating active intrusions using the vulnerability before the patch was even released.

Source: CISA Advisory AA23-325A, "LockBit 3.0 Ransomware Affiliates Exploit CVE-2023-4966 Citrix Bleed Vulnerability", November 2023; Cybersecurity Dive, December 2023

The pattern threat actors followed was not to keep hammering the same vulnerability. They established persistent access, dropped reverse shells or web shells, went quiet, and came back weeks or months later to conduct the actual operation. By the time the ransomware deployed, the initial access vector had long been cleaned up. Most organisations had no idea when or how the attacker got in.

That is what Mythos enables at greater scale and with lower skill requirements. It automates and accelerates the initial access phase.

One important nuance on who moves first. Craig Martin, our Head of IR, notes that standalone threat actor groups, rather than large RaaS operators, are likely to be the first to capitalise on Mythos-class tooling.

Specialised independent groups with the technical capability to integrate new tooling quickly, and without the coordination overhead of a large affiliate network, will be faster to adapt. RaaS operators will follow, but the initial wave is more likely to come from smaller, more agile actors.



Once an attacker is inside, the dwell time problem is compounding. Our DFIR team is seeing dwell times increase. Threat actors are getting in, remaining undetected for weeks, moving laterally, and exfiltrating data.

A large part of why they go undetected is the absence of identity access management and data loss prevention controls that would surface anomalous behaviour before it becomes catastrophic.

The next iteration that genuinely concerns me is what happens when Mythos-class tooling is used not just to get in, but to move through the environment post-compromise.

A model capable of finding subtle vulnerabilities in a 17-year-old kernel codebase is also capable of mapping a network, identifying high-value targets, and generating evasion logic tailored to a specific environment's detection stack.

That is not the present reality. But it is the direction, and organisations that are not building toward lateral movement detection now will be in serious trouble when it arrives.



The Economics of Ransomware Are About to Shift

The ransomware data tells an interesting story right now. According to Chainalysis, victims paid approximately \$813 million in ransomware payments in 2024, a 35 percent decline from the record \$1.25 billion in 2023. The payment rate fell to a historic low. Coveware reported that only 25 percent of victims paid in Q4 2024, down from over 78 percent in 2022.

Source: Chainalysis, "Crypto Crime Ransomware 2025"; chainalysis.com, February 2025

Our own DFIR team has observed the same pattern. Attacks are more targeted, ransom demands are higher, and more organisations are refusing to pay. The model has fewer victims, bigger demands, and lower payment rates.

Mythos disrupts that model. If AI-assisted tooling allows ransomware groups to identify and compromise targets at a dramatically higher volume with lower operational cost, the commercial incentive shifts. The rational play becomes lower demands against more targets, optimised for payment. A hundred organisations paying \$50,000 each generates more revenue than three organisations refusing to pay \$5 million each.

We are already seeing signals of this shift. The 2025 data shows 69 percent of Coveware's cases involved organisations with fewer than 1,000 employees. The mid-market is increasingly in scope. And the average price of initial access broker credentials on dark web markets dropped from \$1,427 in Q1 2023 to \$439 in Q1 2026, driven by automation and AI-assisted tooling.

Source: Darkweb IQ data, cited in sosransomware.com analysis, March 2026

The cost of getting into your network is dropping. The volume of attempts is increasing. The mid-market is now in scope.

The organisations that have historically assumed they are not an interesting enough target should revisit that assumption now.



The cost of getting into your network is dropping. The volume of attempts is increasing. The mid-market is now in scope.

Board Reporting Needs to Evolve

One of the downstream consequences of Mythos that does not get enough airtime is what it demands of board and executive reporting. The metrics most organisations are reporting to their boards today were designed for a different threat environment. Phishing click rates, awareness training completion percentages, and generic incident counts do not give a board what it needs to make risk decisions when exploitation timelines are collapsing.

Here is what good board reporting looks like in this environment. An externally facing Windows system has a critical vulnerability unpatched for 47 days. Taking it offline to patch will cost a material amount in downtime and recovery, depending on how long the maintenance window needs to be and what that system supports.

If it is compromised via a Mythos-identified exploit, the cost is higher still. That is a board-level risk decision, and the board needs the information to make it correctly.

Audit and risk committees should have these specific, named, cost-quantified conversations. Not reviewing hygiene metrics in aggregate. The question is what are the known exposed systems, what is the cost of patching versus the cost of compromise, and who has made the decision.

There is a related issue that belongs in this conversation and rarely gets the board airtime it deserves: technical debt. We see it constantly in our DFIR engagements.

End-of-life systems, unsupported applications, and deprecated services that remain in production because the cost of replacement has been deferred year after year. These systems cannot be patched.

When a Mythos-identified zero-day targets software that is no longer receiving security updates, the organisation has no response option short of taking the system offline entirely. The extent of technical debt across an organisation is a measurable, reportable control failure. It belongs in front of the board as a risk metric, not buried in a technical backlog.

Reporting also needs to be dynamic. The threat landscape does not stand still and your metrics should not either. Working with your MDR provider to



continuously evolve what you report, what detections are in place, and what new threats have emerged is not optional. It should be a standing agenda item in every review.

The Controls That Actually Matter

Nothing that has ever been true about defending an environment has become less true because of Mythos. What has changed is the margin for error. The mindset shift required is from 'if we are compromised' to 'when we are compromised', and the controls that matter most are the ones that limit what an attacker can do after they are already inside.

The areas we are most focused on advising clients right now:

1. **Edge device patching:** This is the highest priority attack surface for Mythos-assisted exploitation. Patch appliances, VPN concentrators, and remote access gateways first and fast. If a device cannot be patched because it is end-of-life, that is a risk decision that needs to be made explicitly, not deferred.
2. **Network segmentation and DMZ architecture:** Assume the perimeter will be breached. Segmentation limits what an attacker can reach once they are inside. Critical systems and high-value data should sit behind controls that require additional authentication to access, so that a compromised edge device does not translate directly into a compromised environment.
3. **Identity access management and unstructured data protection:** IAM controls are critical for detecting lateral movement, but their effectiveness depends on how well you have governed access to unstructured data. File servers, SharePoint libraries, email archives, and shared drives are where most sensitive data actually lives in mid-market organisations. If users have broad access to unstructured data and there is no DLP policy governing how that data can move, an attacker with a single valid credential can exfiltrate significant volumes before any alert fires. Map your unstructured data, classify it, and ensure access is governed by least privilege.
4. **Internal patching, not just perimeter patching:** Patching the edge is essential, but internal systems matter too. Mythos-assisted lateral



movement will target internal vulnerabilities once the perimeter is crossed. An internally-facing system with a known, unpatched vulnerability is an early warning opportunity if you have the visibility to detect exploitation attempts, and a liability if you do not.

5. **Detection and response capability:** You need visibility into what is happening inside your environment, not just at the perimeter. An MDR provider with explicit playbooks for critical advisory response is not optional anymore. That includes detection logic for AI-assisted attack patterns, which look different from traditional exploitation signatures.
6. **Continuous breach attack simulation:** Annual penetration testing is a point-in-time exercise constrained by the skill and time of the tester. BAS driven by curated, industry-specific CTI, feeding into detection engineering and hypothesis-based threat hunting, and closing the loop back to BAS validation, is the assurance model that works in this environment.
7. **Technical debt remediation:** Make end-of-life systems and unsupported applications a board-level metric. Quantify the exposure, prioritise remediation, and where remediation is not feasible, implement compensating controls and make the residual risk visible.

Microsoft's statement from the Glasswing launch is worth sitting with. 'It is clear that these models need to be in the hands of open source owners and defenders everywhere to find and fix these vulnerabilities before attackers get access. Perhaps even more important, everyone needs to prepare for AI-assisted attackers.'

Source: Microsoft, quoted in Anthropic Project Glasswing launch materials, April 2026

Is the Next WannaCry Coming?

It might. Nobody can say definitively. There are credible indications that AI-assisted tooling is being evaluated against Windows system vulnerabilities, and that exploitation research is further along than the public record suggests.

Whether that produces a WannaCry-scale event depends on factors we cannot fully assess.

What we do know is that WannaCry killed organisations that had not patched EternalBlue. The organisations that had applied the MS17-010 patch did not fall. Same logic applies here. Mythos does not overcome good security fundamentals. It punishes the absence of them, faster and at greater volume.

It is also worth being clear that the good side is not standing still. Project Glasswing exists specifically to put these capabilities into the hands of defenders before attackers get equivalent access.

Triskele Labs is already leveraging AI across our platform, including detection engineering, threat intelligence workflows, and our broader MDR capability. The same model capabilities that make Mythos dangerous for finding vulnerabilities make them powerful for building detections, identifying anomalous behaviour, and accelerating incident response.

Mythos is real. The specific nightmare scenario most organisations are imagining is not the immediate threat. The immediate threat is the acceleration of patterns we have been watching for years. Edge exploitation, persistent access, lateral movement, and data theft. Those patterns are well understood. The defences are known.

The question is whether organisations have actually implemented them.

What to Do Now?

If there is one takeaway from this commentary, it is that the patch response question is now existential. Here is the practical list:

1. Map your external-facing systems today. Classify them Red, Amber, Green by criticality. Know which systems can and cannot be taken offline, and at what cost, and for how long.



2. Establish named patch authority. Someone needs to be able to approve an emergency patch outside business hours without a committee. That decision should be made now, not during an incident.
3. Create standing change requests for critical systems so that emergency patching does not require a multi-day approval cycle.
4. Audit your technical debt. Produce a board-visible register of end-of-life and unsupported systems. Prioritise remediation or put compensating controls in place and make the residual risk explicit.
5. Invest in network segmentation. Treat perimeter breach as a when, not an if. Limit what an attacker can reach once they are inside.
6. Prioritise IAM and DLP investment, with specific focus on unstructured data. Map access to file systems, email archives, and collaboration platforms. Apply least privilege. Govern data movement.
7. Move from annual pen testing to ongoing BAS, targeted to current CTI. Test whether your detections actually work, on a cadence that reflects how fast the threat landscape moves.
8. Audit your board reporting. If your metrics have not changed in 12 months, they are not reflecting the current threat environment. Build in a standing review with your MDR provider.
9. Ensure your MDR provider has a documented rapid response playbook for critical advisories. Detection and alerting is not enough. You need a defined response workflow that can operate within hours, not days.

Mythos is a watershed moment. Anthropic has said so themselves, and the technical evidence backs it up. But watershed moments do not necessarily require a new playbook. They require executing the existing playbook better, faster, and with less tolerance for gaps.

The organisations that have done the foundational work are not going to panic when the next headline drops. They are going to be ready for it.





Triskele Labs

www.triskelelabs.com

1300 24 CYBER

Level 16 Queen & Collins Tower

380 Collins St, Melbourne VIC Australia