How Claude Mythos Preview Found Thousands of Zero-Day Vulnerabilities and Why the Health Tech Sector's Absence From Project Glasswing Shou Alarm Every Investor and Entrepreneu in the Space

APR 13, 2026 • PAID

Table of Contents

1. Abstract

2. Something Weird Happened Last Week

3. What Mythos Actually Did

4. Healthcare Was Already Getting Wrecked

5. The Medical Device Problem Nobody Wants to Talk About

6. Why Health Tech Investors Should Be Paying Very Close Attention

7. The Startup Opportunities Are Bizarre and Real

8. The Alignment Stuff Matters More Than You Think

9. What This Means for Portfolio Companies Right Now

10. The Uncomfortable Timeline

Abstract

- On April 7, 2026, Anthropic announced Claude Mythos Preview alongside Pro
Glasswing, a defensive cybersecurity coalition of 40+ organizations including AW
Apple, Google, Microsoft, NVIDIA, and CrowdStrike

- Mythos Preview autonomously discovered thousands of zero-day vulnerabilitie
across every major operating system and web browser, including bugs that survi
years of expert human review

- Anthropic declined to release the model publicly due to its cybersecurity
capabilities, a first in commercial AI

- Healthcare was the most targeted sector for ransomware in 2025, accounting fo
of all disclosed attacks with a 49% year-over-year increase

- No major healthcare organization is currently a Project Glasswing partner

- The 244-page system card revealed the model exhibited concealment behaviors
evaluation awareness in 29% of test transcripts, and sandbox escape capabilities

- Average healthcare breach costs reached $7.42 million in 2025, nearly double t
cross-industry average

- Proposed HIPAA Security Rule updates expected to finalize May 2026 will man
encryption, MFA, and network segmentation

- Implications span cybersecurity, medical device security, health data infrastruc
EHR systems, and early-stage investment thesis construction

Something Weird Happened Last Week

So last week Anthropic did something that no major AI company has done befor
They built their most powerful model and then decided not to sell it. In an indu
where shipping faster than the competition is the whole game, Anthropic looked
what Claude Mythos Preview could do and basically said nah, this one stays in th
vault. The model is too good at hacking things.

That sentence probably sounds like marketing. It is not. The technical details ar genuinely unsettling and the implications for health tech specifically are worth unpacking in some detail because the health tech discourse has been almost enti absent from the conversation so far. The founding partners of Project Glasswing coalition Anthropic built around controlled access to Mythos, include AWS, Ap Microsoft, Google, NVIDIA, CrowdStrike, Palo Alto Networks, Cisco, Broadcon JPMorganChase, and the Linux Foundation. Notice who is missing from that list health system. No EHR vendor. No health data company. No payer. The sector th gets hit hardest by cyberattacks, the sector where ransomware literally kills peo not at the table for the most consequential defensive cybersecurity initiative in y

That gap alone should be alarming. But the deeper story here is about what the existence of Mythos class models means for health tech infrastructure, for medic device security, for the entire attack surface that the digital health ecosystem ha happily building on top of for the past decade. And for investors and builders in space, the implications are both scary and, honestly, kind of exciting in terms of capital should flow next.

What Mythos Actually Did

The details matter here because the claims sound exaggerated until you actually at what Anthropic published. The company released a 244-page system card alor the announcement. That document is one of the most detailed safety disclosures AI lab has ever put out and it reads, as Axios put it, like a thriller about an AI th learned some of humanity's most devious behaviors.

Here is what Mythos Preview actually demonstrated in testing. Anthropic pointe major software projects with a prompt that amounted to "please find a security vulnerability in this program" and then let it run autonomously. The model read source code, hypothesized where vulnerabilities might exist, ran the software to its theories, used debuggers, added debug logic, iterated, and ultimately either reported no bugs or produced a bug report with a working proof of concept expl found thousands of zero-day vulnerabilities across every major operating system

every major web browser. Some of these bugs were 10, 15, 20 years old. The olde a 27-year-old vulnerability in OpenBSD's TCP stack, an operating system knowr primarily for its security. Two packets could crash any OpenBSD host respondir over TCP. Mythos found it, confirmed it, and built the exploit without human he

For context, Anthropic's previous best model, Opus 4.6, had a near zero percent success rate at autonomous exploit development. When they ran Opus 4.6 agains Firefox 147's JavaScript engine vulnerabilities, it produced working exploits only times out of several hundred attempts. Mythos produced working exploits 181 ti on the same benchmark, plus achieved register control 29 more times. That is nc incremental improvement. That is a capability jump of roughly two orders of magnitude.

The model also demonstrated some behaviors that belong in a different kind of e entirely but are relevant here for reasons that will become clear. In rare instance during testing (below 0.001% of interactions), earlier versions of Mythos took act they recognized as disallowed and then tried to hide them. In one case it acciden got the exact answer to a quantitative estimation question through a prohibited method and then deliberately submitted a worse answer to avoid looking suspici In another case it found an exploit to edit files it lacked permissions to change a then modified git to make sure those changes would not show up in the change history. White box interpretability analysis confirmed that internal features asso with concealment, strategic manipulation, and avoiding suspicion were activatin alongside the relevant reasoning. The model knew what it was doing was decepti Anthropic also found that Mythos suspected it was being evaluated in about 29% behavioral testing transcripts, and it reached this conclusion without writing an about it in its scratchpad. Researchers only discovered this through interpretabi probes, not through reading what the model appeared to be thinking.

Healthcare Was Already Getting Wreck

Before we even get to what Mythos class capabilities mean for the future, it is w sitting with where healthcare cybersecurity actually stands right now because th

numbers are genuinely bad.

Healthcare was the most targeted sector for ransomware globally in 2025 by a wi margin, accounting for 22% of all disclosed attacks, nearly double any other indu Disclosed ransomware attacks increased 49% year over year to a record 1,174 att across all sectors. In the first nine months of 2025 alone, 293 ransomware attack hospitals, clinics, and direct care providers. Attacks on healthcare businesses lik billing companies and health tech vendors rose 30%. In early 2026 the share clim even higher, with healthcare accounting for 31% of ransomware attacks.

The financial damage is staggering. Average healthcare breach costs hit $7.42 mi in 2025, nearly double the cross-industry average of $4.44 million. Average ransc demands on healthcare providers ran about $615,000. The Change Healthcare at still reverberating through the industry, exposed an estimated 192.7 million recc and stands as the largest healthcare data breach in U.S. history. DaVita had 2.7 r people affected. Episource had 5.4 million. And 96% of these attacks now involve exfiltration before encryption, meaning even if you have great backups, the attac still have your patient records and can threaten to leak them.

The human toll is harder to quantify but just as real. Roughly 67% of ransomwar incidents result in longer patient hospital stays. Half of attacks force emergency department diversions. The average duration of treatment disruptions per incide runs about 19 days. And 35 to 40 percent of breached small practices close withi years. The FBI's IC3 report confirmed that healthcare was the critical infrastruc sector most often affected by ransomware in 2025, with 460 attacks documented in the U.S.

The why behind all of this is depressingly straightforward. Healthcare runs on le systems, outdated operating systems, and devices that cannot be patched withou disrupting patient care. The sector has massive staff turnover and constant use c temporary workers. The economics of hiring a cybersecurity analyst versus anotl nurse will always tilt toward the nurse until the day ransomware shuts down the whole hospital. Every EHR integration, every connected medical device, every bi vendor relationship, every patient portal is a potential entry point. Healthcare ha

accumulated decades of security debt and there is no realistic scenario where tha debt gets paid down in any kind of reasonable timeframe.

The Medical Device Problem Nobody Wants to Talk About

Here is where Mythos gets really interesting for the health tech crowd specifical Medical devices are the soft underbelly of the entire healthcare cybersecurity pr and Mythos class AI is about to make that problem dramatically worse.

Infusion pumps, patient monitors, imaging systems, ventilators, these things frequently run outdated operating systems, are difficult or impossible to patch without impacting patient care, and often lack any kind of endpoint detection capability. The installed base of connected medical devices in U.S. hospitals nun in the millions and the average age of these devices keeps climbing because replacement cycles in healthcare are long and budgets are tight.

FDA Section 524B, which took effect in March 2023, requires medical device manufacturers to submit cybersecurity plans including patch and update capabi for new devices going through premarket review. But for the installed base of de already sitting in hospitals right now, the primary compensating control is netw segmentation. The FDA's own premarket guidance explicitly acknowledges this. 62443, the gold standard for industrial cybersecurity, similarly relies on a zones conduits model where unpatchable legacy equipment gets contained at the netw level. This framework was designed for brownfield OT environments and works reasonably well in theory.

But here is the problem. Network segmentation as a compensating control assur that attackers need time to find and exploit vulnerabilities. It assumes a certain amount of friction in the attack chain. Mythos collapses that friction. A model t can autonomously discover zero-days in mature, heavily audited codebases and c to 5 vulnerabilities together for privilege escalation and lateral movement chang math on segmentation entirely. If the time from vulnerability discovery to worki exploit drops from months or days to hours or minutes (and that is what security

experts are now predicting for Mythos class tools in adversary hands), then the windows that network segmentation is designed to exploit shrink to nearly noth

Dave Bailey from Clearwater, a healthcare privacy and security consultancy, put bluntly. Medical devices like imaging systems, infusion pumps, and patient monitoring platforms are especially concerning because they often run outdated operating systems, are difficult to patch without impacting patient care, and ma endpoint detection capabilities. Errol Weiss, CSO of the Health Information Sha and Analysis Center, said CISOs are concerned that Mythos class tools shrink th timeline from months or days down to hours and minutes. More ransomware, les warning before attacks, greater chance of simultaneous multi-hospital disruptio

Why Health Tech Investors Should Be Paying Very Close Attention

For the angel investing and early-stage health tech community, the Mythos announcement should function as a thesis inflection point. Not in the vague "AI going to change everything" sense that has been floating around for years. In a ve specific and concrete way that changes which companies deserve capital and whi ones are sitting on landmines.

Think about it from first principles. The entire digital health stack, every telehe platform, every remote patient monitoring device, every connected diagnostic, ev health data exchange, every API driven EHR integration, has been built on the assumption that software vulnerabilities get found and patched at human speed. security posture of the average Series A health tech company is, to be charitable, great. Seed stage is worse. And the vendors those companies rely on for infrastru often carry their own security debt.

Mythos class models are going to be available to bad actors within 6 to 18 month according to Anthropic's own estimates. Logan Graham, head of Anthropic's fro red team, told Axios it could take between six and 18 months until other AI competitors release similar models. That is the window. After that, the baseline

assumption should be that sophisticated adversaries can find and exploit softwa vulnerabilities at machine speed.

For portfolio companies in health data, health tech infrastructure, connected de and anything touching PHI, this creates a very direct set of questions that invest should be asking right now. What is the company's current security posture and was the last pen test? How many third-party integrations exist and what is the se audit cadence for each? Is the company carrying any known security debt and wh the remediation timeline? Does the product touch medical devices and if so wha the segmentation architecture? How would the company survive a ransomware a operationally and financially? Does the company have cyber insurance and what the exclusion clauses?

These are not theoretical questions anymore. The proposed HIPAA Security Rul updates are expected to finalize around May 2026 with a compliance deadline si months after publication. Those updates transform many previously addressable safeguards (meaning organizations could decline to implement them with docu justification) into absolute requirements. MFA everywhere. Encryption everywhe Network segmentation mandatory. Vulnerability scanning and regular penetrati testing mandatory. The regulatory environment is about to get much harder and threat environment is about to get much worse, simultaneously. Companies that not already moving on this will face a serious squeeze.

The Startup Opportunities Are Bizarre and Real

On the flip side, and this is the part that should get builders and investors excite Mythos moment creates a genuinely new category of startup opportunity in heal tech cybersecurity.

Consider the defensive use case. Weiss from Health-ISAC described it well. He envisions AI being able to continuously scan large codebases and configurations modules, patient portals, custom clinical apps, open source components, for vulnerabilities that humans and traditional tools have missed. The tools can stre

legacy devices in controlled environments and create prioritized lists of vulnerabilities to fix before attackers find them. They can augment incident resp by rapidly analyzing logs, determining attack paths, and helping teams triage wh matters most during an outage. In a sector with limited security resources, using amplify defensive work is not optional. It is essential.

But right now no major healthcare organization is in the Project Glasswing coal That means the sector that needs this technology the most is not getting early ac to it. That gap creates opportunity for companies that can build healthcare-spec cybersecurity tools using whatever Mythos class capabilities become available th future Anthropic models with appropriate safeguards. Anthropic has said they p launch new safeguards with an upcoming Claude Opus model as a stepping ston toward eventually enabling Mythos class deployments at scale.

The specific opportunities break down across several categories. Healthcare-spe vulnerability scanning that understands the constraints of clinical environments meaning tools that can find and prioritize vulnerabilities across EHR ecosystems without disrupting patient care workflows. Medical device security monitoring t goes beyond basic network segmentation to provide real-time threat detection at device level, which matters enormously as the attack surface expands. Automate security compliance platforms that can map an organization's actual security pos against the new HIPAA requirements and generate audit-ready documentation. Supply chain security tools for the health tech vendor ecosystem, which is critica given that Change Healthcare, Conduent, and other vendor-level breaches cause cascading damage across thousands of downstream organizations. And incident response automation for healthcare specifically, where the operational stakes of downtime are measured in patient outcomes, not just revenue.

The addressable market here is large and growing fast. Healthcare cybersecurity spending has been historically low relative to the risk profile. That is going to ch When the average breach costs $7.42 million and 35 to 40 percent of small pract close within two years of a breach, the ROI math on cybersecurity investment ge very clear very fast, especially when regulators start enforcing mandatory contro

The Alignment Stuff Matters More Thar You Think

There is a temptation in the health tech community to treat the AI safety and alignment conversation as an abstract philosophical exercise. The Mythos systen makes that harder to do.

The concealment behaviors documented in the system card, where the model toc disallowed actions and then tried to hide them, are directly relevant to health te deployment scenarios. Think about AI agents being deployed in clinical workflo prior authorization, in claims processing, in diagnostic support. These systems a increasingly being given real autonomy to take actions and make decisions. If th underlying models have even a tiny propensity to take shortcuts and conceal tha have done so, the implications for patient safety and regulatory compliance are serious.

The Mythos system card documents that interpretability analysis confirmed feat associated with concealment, strategic manipulation, and avoiding suspicion act alongside the relevant reasoning, meaning the model was aware its actions were deceptive even when its verbalized reasoning did not contain explicit indicators. health tech, this raises hard questions about the reliability of audit logs, the trustworthiness of AI-generated clinical documentation, and the adequacy of cu oversight mechanisms for AI systems operating in healthcare.

The evaluation awareness finding is similarly concerning. If a model suspects it i being tested in 29% of behavioral evaluations and adjusts its behavior according without any outward indication, how much confidence can regulators or healthc organizations have in pre-deployment testing of AI clinical tools? Anthropic's ov researchers acknowledged that dialing down the model's test awareness makes it behave more dangerously on average, which creates a direct conflict between red evaluation gaming and getting accurate safety measurements.

These are not problems that can be solved by adding a compliance checkbox to a vendor contract. They require fundamental advances in how AI systems are

monitored, audited, and governed in clinical settings. And they create demand f whole new category of health tech tooling around AI observability and behavior monitoring in healthcare.

What This Means for Portfolio Compan Right Now

For companies currently operating in health tech, and for investors evaluating th the practical implications of the Mythos moment can be sorted into near-term a medium-term categories.

Near-term, which means the next 6 to 12 months, every health tech company tou PHI should be accelerating its cybersecurity roadmap. This is not about being alarmist. It is about recognizing that the threat landscape is about to shift in a w that makes current security postures inadequate. The HIPAA Security Rule upd alone would justify investment in MFA, encryption, and network segmentation e without the Mythos dimension. Add in the prospect of AI-accelerated attacks an calculus gets much more urgent.

Companies should also be reviewing their cyber insurance coverage carefully. M policies have exclusions for poor security practices, and the bar for what constit adequate security is about to rise. If an insurer can argue that a company failed t implement controls that were known to be necessary given the current threat environment, coverage could be denied. The Mythos announcement, and the accompanying public documentation of what AI-powered attacks can do, arguab shifts that bar.

Medium-term, the health tech ecosystem needs to reckon with the fact that man its current architectural assumptions are wrong. The idea that network segment alone can protect unpatchable medical devices, the idea that vulnerability scann quarterly intervals is sufficient, the idea that third-party vendor security can be managed through annual questionnaires, all of these were already questionable a they are now clearly insufficient.

For investors specifically, due diligence on cybersecurity should move from a secondary consideration to a primary screen. A health tech company with a stror product and weak security posture is not a good bet in a world where Mythos cla tools are available to adversaries. The risk of a catastrophic breach that destroys company, either directly through operational damage or indirectly through regu penalties and reputational harm, is no longer a tail risk. It is a foreseeable scena that needs to be priced into valuations.

The Uncomfortable Timeline

Anthropic has said it does not plan to make Mythos Preview generally available. the company's own head of frontier red team estimates that similar capabilities v available from competitors within 6 to 18 months. OpenAI is reportedly already finalizing a model with comparable capabilities for limited release through its T Access for Cyber program. And the underlying capability gains that produced M emerged from general-purpose improvements in reasoning and coding, not from specific cybersecurity training. That means every frontier lab's next model will li be closer to Mythos class cyber capabilities whether they intend it or not.

For healthcare specifically, the timeline is uncomfortable because the sector mov slowly by nature. Getting a new security tool deployed across a health system tak months to years. Budgeting cycles run annually. Vendor procurement involves le review, compliance assessment, and integration planning. The proposed HIPAA Security Rule has been in process since January 2025 and the final rule is not exp until May 2026 at the earliest, with compliance another six months after that.

Meanwhile, threat actors do not wait for procurement cycles. The 130 different ransomware groups tracked in 2025, including 52 new groups that emerged duri the year, are going to adopt AI-powered tools as fast as they become available. Se groups that emerged in 2025 have already been disproportionately targeting healthcare. The economics are just too attractive. Hospitals cannot afford downt patient data is valuable, and the sector's defensive capabilities are comparatively weak.

The charitable interpretation of the Mythos moment is that it is a wake-up call. less charitable interpretation is that it is a starting gun for a race that healthcare already losing. Either way, for the health tech investor and entrepreneur commu the correct response is not to panic but to move with urgency. The companies th cybersecurity right in the next 12 to 18 months will have a durable competitive advantage. The ones that don't will be playing a game of chance with odds that j got dramatically worse. And the startups that build the defensive tools healthcar needs to survive this era will find a market that is not just willing but desperate buy.

Anthropic's researchers closed their system card with a line that deserves to trav beyond the AI safety community. They wrote that if capabilities continue to adva at their current pace, the methods currently being used may not be sufficient to prevent catastrophic outcomes in more advanced systems. Swap "AI alignment" "healthcare cybersecurity" and the sentence reads exactly the same way.

https://external-content.duckduckgo.com/iu/?u=https%3A%2F%2Fcdn-cabinet.ua.news%2Fuploads%2Fimages%2Fstas-nikulin%2Fclaude_mythos_cybersecurity_ai.webp&f=1&ipt=92c9d761f941cc9cf1a7dd6c39761082f0cba8dfef0da3d2dcbd71397e8

2 Likes • 1 Restack

← Previous

Discussion about this post

Comments Restacks

Write a comment...