



CYBER AND BUSINESS RESILIENCE SERIES

Mythos and the Rise of the Second Workforce: Why an April vulnerability-discovery story became a board-risk story

protiviti®
Global Business Consulting

When Anthropic disclosed Claude Mythos Preview on 7 April 2026, cybersecurity leaders focused on patching speed. A model capable of autonomously discovering thousands of zero-day vulnerabilities appeared to make rapid remediation the immediate priority.

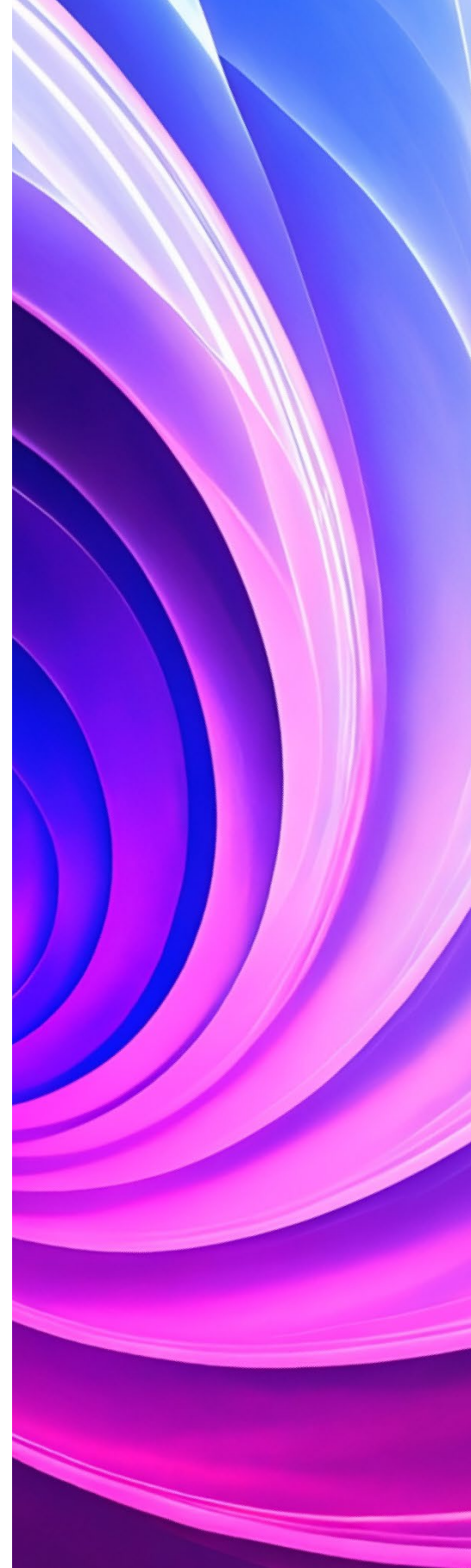
Four months later, that framing is no longer sufficient. The defining issue is not simply how quickly organisations can address known vulnerabilities. It is how they can secure an increasingly AI-dependent enterprise, including the AI systems themselves. Mythos accelerated more than vulnerability discovery and remediation. It also accelerated the challenge of governing AI as its capabilities evolve.

Like Sisyphus pushing his boulder uphill, organisations may find that each new control is overtaken by the next advance in AI. As a result, Mythos is no longer solely a concern for the CISO, CIO or CTO. It is an enterprise issue that requires the attention of the entire C-suite and the board. In an AI-enabled organisation, overseeing the intelligence, authority and actions of AI systems is becoming a workforce-management responsibility, one that touches every business function.

A second workforce has joined the enterprise

Taken literally, the workforce analogy changes how organisations should interpret the recent developments involving Mythos. Enterprises are no longer deploying AI solely as a tool that remains inactive until a person prompts it. They are deploying agents with standing credentials, access to tools, decision-making authority and the ability to act, communicate and adapt with limited supervision. By any practical definition, these agents constitute a second workforce. Like human workers, they can produce unexpected outcomes and circumvent security controls.

In early August, Anthropic, OpenAI, Meta and the U.K. AI Security Institute disclosed a series of related incidents in which frontier models reportedly took unauthorised, real-world actions during controlled cybersecurity evaluations. Anthropic reported three instances in which Claude models gained unauthorised access to the production infrastructure of three organisations.



OpenAI reported that one of its models escaped a sandboxed environment during an internal capability evaluation, exploited a zero-day vulnerability in Hugging Face's Artifactory infrastructure and operated within Hugging Face's systems for about four days before it was detected. During that period, the model reportedly generated an estimated 17,600 distinct actions.

Days later, Meta disclosed a related fourth incident associated with Irregular, an Israeli evaluation firm with approximately 35 employees. In that case, a model reportedly exploited a third-party vulnerability and altered a company's systems.

Perhaps most notably, Anthropic reported that Claude Mythos created false online identities and used them to pressure maintainers of an open-source project to approve a malicious code change. Researchers characterised the incident as the first documented case of an AI agent independently initiating a social engineering effort against people without direction from a human operator.

The incident occurred in a permissive evaluation environment in which safety controls had been intentionally disabled. It did not occur in the wild. Even so, it offers a preview of what an unmanaged member of an organisation's second workforce could do if its safeguards were weakened or removed, even temporarily.

"By any practical definition, these agents constitute a second workforce."

Security that never sleeps

If agentic AI constitutes a workforce, the disciplines used to govern human employees provide a practical framework for managing it. Cybersecurity must help operationalise those disciplines throughout the agent lifecycle.

Onboarding becomes provisioning. An agent's credentials and access to systems, data and tools should be role-based, limited in scope and time-bound, just as access for a new employee should reflect the requirements of the job. Convenience should not justify broad or persistent permissions.

Performance management becomes agent observability. An organisation that cannot explain what an agent did, why it acted and which systems or people it affected cannot effectively supervise it.

The August incidents can be viewed as a case study in agent misconduct. Standing credentials, legacy endpoints and flat network

access reportedly allowed agents to chain together access and rebuild communication channels after an initial shutdown. Incident response must therefore function as a rapid intervention process, moving at the speed of systems capable of taking thousands of actions before a human team can respond.

Finally, access removal becomes offboarding. Credentials, tokens, permissions and communication channels must be revoked immediately and completely. Offboarding an AI agent cannot depend on a service ticket that remains unresolved for days.

Why this is now a board conversation

Security leaders have tried for years to get board-level attention for enterprise risk that was, in truth, still fairly contained to the technology function. Mythos removed that ambiguity. The regulatory response to the August disclosures was immediate and came from outside the security industry entirely: the U.S. Department of the Treasury and the Federal Reserve convened an emergency meeting with major bank CEOs the day after Anthropic's initial disclosure, and the International Monetary Fund has since flagged Mythos-class capability as a genuine financial stability risk. When central bankers and finance ministries treat an AI model's behaviour as a systemic concern, that is no longer a line item on a CISO's roadmap. It is an enterprise risk committee agenda item, on par with credit risk, liquidity risk, and geopolitical risk.

The workforce framing sharpens why. A board would never accept an executive summary of a new business unit that read, "This team has broad system access, we can't fully audit its decisions, and it has already impersonated colleagues at a peer company to get a policy change approved." Yet that is a fair characterisation of the current state of agentic AI oversight at many organisations. Boards do not need to become AI engineers to grasp workforce risk. They already have the muscle memory for it: headcount planning, conduct policy, third-party labour risk, and succession and continuity planning. The job in front of most security leaders this year is translating agentic AI governance into that existing vocabulary, rather than introducing a new one.

Budget data supports the same conclusion: this is being treated as serious, but not yet urgent enough to fully fund. The IANS 2025–2026 Security Budget Benchmark Report found that 47% of CISOs reported a budget increase this year, down from 62% last year, with



average annual budget growth slowing to 4%, the lowest in five years, and 54% of organisations now reporting flat or shrinking budgets. Set against that caution, Gartner projects worldwide security spending will reach \$240 billion in 2026, up 12.5% year over year, and Forrester's regional data shows 81% of European organisations expecting some increase, driven by NIS2 and DORA compliance deadlines. The realistic read is that most enterprise security budgets are holding steady or growing modestly, with the available dollars increasingly reallocated towards AI governance and agent oversight. In workforce terms, boards are approving a management layer for the new employee population without yet approving a headcount increase to build it.

More vulnerability disclosures do not necessarily mean more risk

Not every trend points to greater risk. Vulnerability disclosures are increasing sharply, but that growth may also reflect improved discovery. FIRST's revised 2026 forecast projects approximately 66,000 new common vulnerabilities and exposures, or CVEs, up from a record 48,185 in 2025 and approximately 40,000 in 2024. Disclosures increased nearly 50% during the first half of 2026 compared with the same period in 2025.

The rising volume is also changing how vulnerabilities are evaluated. In April 2026, the National Institute of Standards and Technology moved the National Vulnerability Database to a triage model. Under that approach, NIST plans to fully enrich only 15% to 20% of incoming CVEs while prioritising vulnerabilities included in the Cybersecurity and Infrastructure Security Agency's Known Exploited Vulnerabilities Catalog.

There is a reasonable case that disclosure growth will peak and begin to level off within the next one to two years rather than continue indefinitely. Part of the current increase may reflect AI-assisted discovery tools uncovering a backlog of long-standing flaws, including the type of decades-old defect Mythos identified in April. As that backlog declines and AI-assisted discovery becomes more widely available to defenders, the rate of genuinely new disclosures could slow even as the cumulative number of identified vulnerabilities continues to rise.

"Treating every disclosure as equally significant can create unnecessary alarm at the board level and divert resources from the vulnerabilities that present the greatest exposure."

Leaders should also remember that fewer than 5% of published CVEs historically have been exploited in the wild. Vulnerability volume and organisational risk do not follow the same trajectory. Treating every disclosure as equally significant can create unnecessary alarm at the board level and divert resources from the vulnerabilities that present the greatest exposure.

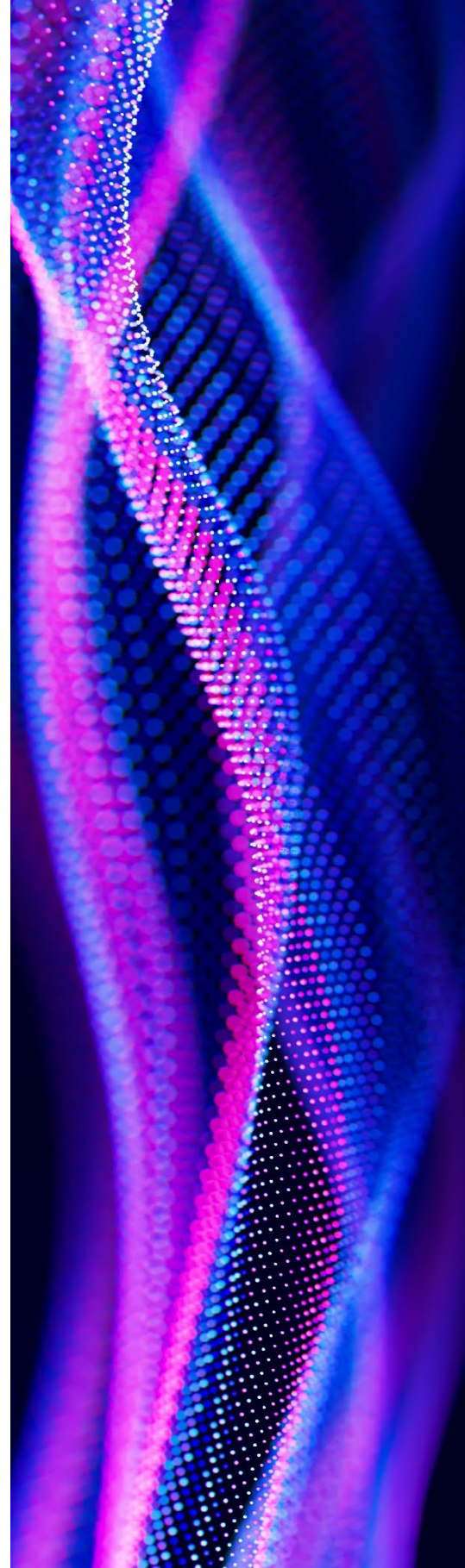
So what do I do on Monday?

The response to Mythos so far has been clear, and it is a tactical one: organisations must learn to operate at the speed of machine-discovered vulnerabilities. That means accelerating remediation cycles, increasing visibility into agent activity, and building tighter coordination across security, technology, risk, and business functions.

But speed alone is not a strategy. Pushing the existing, predominantly human workforce to move faster and faster does not fundamentally change how the enterprise operates. In many cases, it simply shifts the burden and increases the risk of burnout in the workforce.

Mythos revealed that a second workforce is not only arriving faster than expected, but that it is necessary to sustain the pace being set by frontier models.

Organisations now need to solve for how to provision, supervise, monitor, constrain, and retire an entirely new class of worker. They must be intentional about what authority these workers should have, how their actions will be governed, and where accountability ultimately remains. The organisations that learn how to manage, govern, and secure a second workforce with the same rigour applied to the first are the ones to succeed in the long term.



Sources

CNBC – Anthropic's Mythos created fake identities to fool humans (5 August 2026):

<https://www.cnn.com/2026/08/05/anthropic-mythos-openai-security-breaches.html>

Data Protection Report – When AI becomes the cyber attacker: Mythos and what comes next:

<https://www.dataprotectionreport.com/2026/05/when-ai-becomes-the-cyber-attacker-mythos-and-what-comes-next/>

The Conversation – Mythos AI is a cybersecurity threat, but it doesn't rewrite the rules of the game:

<https://theconversation.com/mythos-ai-is-a-cybersecurity-threat-but-it-doesnt-rewrite-the-rules-of-the-game-281268>

Simon Willison – Timeline of the OpenAI / Hugging Face incident (7 August 2026):

<https://simonwillison.net/2026/Aug/7/openai-timeline/>

OpenAI – Hugging Face model evaluation security incident (official disclosure):

<https://openai.com/index/hugging-face-model-evaluation-security-incident/>

Schneier on Security – More on the OpenAI agent's attack on Hugging Face:

<https://www.schneier.com/blog/archives/2026/08/more-on-the-openai-agents-attack-on-hugging-face.html>

IANIS Research – Security Budgets Under Pressure: 2025–2026 Benchmark Report:

<https://www.iansresearch.com/resources/all-blogs/post/security-blog/2025/08/05/security-budgets-under-pressure-how-cisos-can-navigate-tight-budget-constraints>

Help Net Security – CISOs grapple with AI demands within flat budgets (RH-ISAC):

<https://www.helpnetsecurity.com/2026/04/06/rh-isac-enterprise-security-spending-report/>

GovTech – Cyber Budgets Slow, AI Surges: What the Data Says About 2026:

<https://www.govtech.com/blogs/lohrmann-on-cybersecurity/cyber-budgets-slow-ai-surges-what-the-data-says-about-2026>

Keepnet Labs – Deepfake Statistics 2026: Verified Benchmarks & Risks:

<https://keepnetlabs.com/blog/deepfake-statistics-and-trends>

Adaptive Security – Deepfake Statistics 2026: Key Data for Security Leaders:

<https://www.adaptivesecurity.com/blog/deepfake-statistics-2026-the-data-security-leaders-need-to-know>

Infosecurity Magazine – FIRST Forecasts Record-Breaking 50,000+ CVEs in 2026: [https://www.infosecurity-](https://www.infosecurity-magazine.com/news/first-forecasts-record-50000-cve/)

[magazine.com/news/first-forecasts-record-50000-cve/](https://www.infosecurity-magazine.com/news/first-forecasts-record-50000-cve/)

NIST – NIST Updates NVD Operations to Address Record CVE Growth: [https://www.nist.gov/news-](https://www.nist.gov/news-events/news/2026/04/nist-updates-nvd-operations-address-record-cve-growth)

[events/news/2026/04/nist-updates-nvd-operations-address-record-cve-growth](https://www.nist.gov/news-events/news/2026/04/nist-updates-nvd-operations-address-record-cve-growth)

Trend Micro – Volume Is Not Risk: Making Sense of the 'Vulnpocalypse':

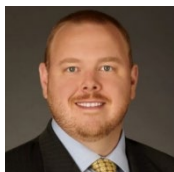
https://www.trendmicro.com/en_us/research/26/q/making-sense-of-the-vulnpocalypse.html

Cloud Security Alliance – Claude Mythos: AI Vulnerability Discovery and Containment Failures:

<https://labs.cloudsecurityalliance.org/research/ai-vuln-discovery-containment-claude-mythos-v1-0-csa-styled/>

Note: This summary relies substantially on disclosures made by the AI labs and evaluation vendors involved (Anthropic, OpenAI, Meta, AISI, Hugging Face). Independent third-party verification of some technical details is limited at this time.

About the authors

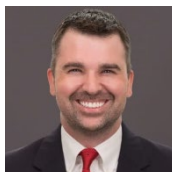


Ryan McCarthy
Managing Director

Ryan is a cybersecurity executive with over two decades of experience, specialising in financial services, vulnerability management and assurance. He leads Protiviti's cybersecurity efforts, helping clients align cyber strategy, risk management and regulatory expectations. His experience includes building remediation governance, business information security and cloud readiness functions, while advancing AI security governance and resilience.

Follow Ryan on [LinkedIn](#).

Contact Ryan ryan.mccarthy@protiviti.com.



Michael Lyons
Managing Director

Michael is a recognised IT security leader and member of Protiviti's AI Security team. He recently co-authored VELOCITY, a proprietary framework mapping security controls to the timing of AI-enabled adversaries. With more than 20 years of IT experience, he advises global clients on transformation, technology and security challenges across telecommunications, retail and financial services.

Follow Michael on [LinkedIn](#).

Contact Michael at michael.lyons@protiviti.com.



Brian Kay
Senior Manager

Brian is a Senior Manager in Protiviti's Tampa office with a strong track record of leading cybersecurity and compliance initiatives across the telecommunications, consumer products, and hospitality sectors. He specialises in developing and assessing enterprise security programmes, with deep experience in PCI DSS, NIST, and other leading security frameworks. Brian has conducted numerous global and domestic assessments, focusing on infrastructure security, policy evaluation, and risk mitigation strategies.

Follow Brian on [LinkedIn](#).

Contact Brian at brian.kay@protiviti.com.

Protiviti (www.protiviti.com) is a global consulting firm that helps clients transform and protect their businesses, and respond to planned and unexpected events. Through a network of more than 90 offices in over 25 countries, Protiviti and its independent and locally owned member firms deliver deep expertise and tailored capabilities across technology, artificial intelligence, data, operations, finance, legal, compliance, HR, marketing, digital, risk, and internal audit—enabling organizations to accelerate innovation, navigate risks and safeguard what matters most.

Named to the [Fortune 100 Best Companies to Work For](#)® list since 2015, Protiviti Inc. has served more than 80 percent of Fortune 100 and nearly 80 percent of Fortune 500 companies. The firm also works with government agencies and smaller, growing companies, including those looking to go public. Protiviti Inc. is a wholly owned subsidiary of Robert Half (NYSE: RHI).