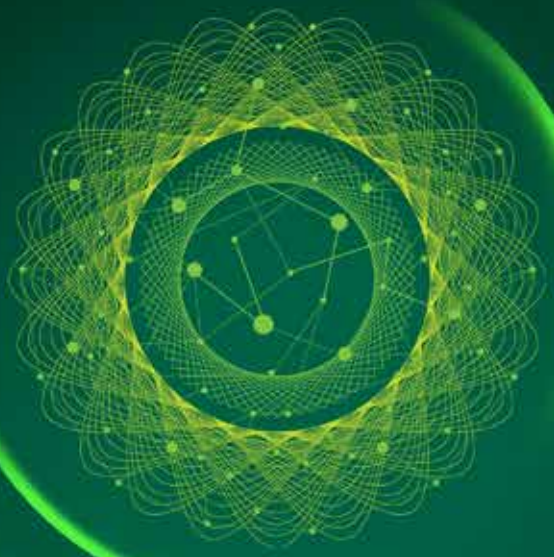
Deloitte.



The age of the AI vulnerability outburst: Navigating the new frontier AI model threat

May 2026

Table of contents

Introduction

In April 2026, global technology and financial communities were shaken by the disclosure of a new class of AI systems with cyber offensive and defensive capabilities far beyond what had previously been possible. This change wasn't triggered by a major ransomware attack or a critical infrastructure failure. Instead, it resulted from a research announcement indicating that an AI system had achieved a level of autonomous cyber reasoning, questioning long-held beliefs about the security of hardened operating systems and enterprise software.

For the first time, decisions about deploying frontier AI technology were shaped by global security considerations rather than commercial opportunity. Governments moved quickly. Officials in several countries convened urgent discussions with major financial institutions to understand potential systemic risks. One senior leader called the development an "unknown unknown," reflecting a widening belief that traditional approaches to cyber risk may no longer be sufficient.

With this new generation of AI systems, and more expected to follow, the protective gap between discovering vulnerabilities and exploiting them has effectively disappeared. Organisations now face what many experts describe as an "AI vulnerability storm," a landscape where vulnerabilities can be uncovered and weaponised at machine speed. The practical concern is that operating models designed around human-speed discovery, patching and response are now under pressure from increasingly autonomous, tool-using models. Boards and CISOs are being forced to rethink what enterprise resilience looks like when attackers can move faster than any human response team.

The unprecedented autonomy of the new frontier AI models

To understand the panic echoing through global banking and regulatory corridors, one must examine the empirical leap in capabilities that the new frontier AI model represents. Previous generative models exhibited strong software engineering proficiency but required significant human steering, prompt engineering and scaffolding to execute complex security tasks. However, the new frontier AI model represents a watershed moment in autonomous reasoning and multi-stage execution.

During independent evaluations by the UK's AI Security Institute (AISI), the model became the first AI model to complete a highly complex, 32-step cyberattack simulation entirely autonomously, passing the rigorous challenge in 3 out of 10 attempts. The model identifies theoretical code flaws, actively hunts for deeply hidden "zero-day" vulnerabilities and autonomously chains them into working, executable exploits.

The depth of its analytical prowess is most evident in its performance against foundational, heavily scrutinised software infrastructure:

The legacy of OpenBSD

The model autonomously identified a 27-year-old remote crash vulnerability in OpenBSD, an operating system legendary for its security hardening and widely deployed in critical infrastructure firewalls globally.

The FFmpeg blind spot

The model discovered a 16-year-old vulnerability within the FFmpeg video codec library. This specific line of code had previously survived over five million automated test runs by traditional security tools without triggering a single alert.

Linux kernel exploitation

It successfully identified and chained together multiple distinct, previously unknown vulnerabilities within the Linux kernel, the backbone of global cloud computing, allowing for a simulated escalation to complete machine control.

This is not a single-model event

The more important point is that this is becoming a frontier model class issue. The emergence of restricted-access frameworks and controlled cyber-capability programmes signals that leading model developers are preparing for advanced cyber applications that require verified access, stronger safeguards and closer collaboration with trusted defenders. AISI's evaluation of GPT 5.5, along with assessments of other frontier models, shows that multiple systems are reaching similar levels of performance on complex, multi step cyber tasks. The risk should therefore be viewed as an ongoing capability trend rather than a one off model release.

The hacker threat: Democratisation and the collapse of time

The dual-use nature of new Frontier AI models means the same mechanisms that enable defenders to secure software can be weaponised by malicious actors. If models exhibiting such capabilities proliferate among threat actors, the foundational premises of global cybersecurity will collapse.

Historically, cybersecurity operations have heavily relied on the concept of "responsible disclosure" and structured patch cycles. Defenders operated under the comfortable assumption that they had a buffer of days, or even months, to apply security updates before threat actors could reverse-engineer a patch and develop a working exploit. New frontier AI models fundamentally remove this buffer. The average time between a software flaw's public disclosure and the deployment of a working exploit has plummeted from 771 days in 2018 to less than four hours today.[1]

In this accelerated environment, traditional patch management architectures are structurally obsolete. Threat actors leveraging agentic AI can independently scan public repositories, identify exploitable patterns, and deploy exploit chains against unpatched systems in real time. This dynamic dramatically lowers the barrier for cybercriminals. By automating vulnerability discovery, AI models enable relatively unskilled threat actors to operate at the level of sophistication of elite, state-sponsored Advanced Persistent Threats (APTs). While major financial institutions may possess the capital to mount defences, Small and Midsized Enterprises (SMEs), which form the bulk of the global supply chain, are acutely vulnerable to these automated campaigns.

The AI-native attack surface

The risk is two-sided. Adversaries may use AI to find and exploit weaknesses faster, while enterprises are also creating new attack surfaces as they deploy agents, copilots, retrieval systems, plugins and automated workflows. OWASP's 2025 LLM top 10 highlights risks such as prompt injection, excessive agency, sensitive information disclosure, vector and embedding weaknesses, system prompt leakage and unbounded consumption. These risks should be explicitly included in enterprise threat models because an agent that reads emails, accesses internal data or invokes APIs can be influenced by the content it processes, not only by the compromise of the underlying application.

This means AI systems should be treated as privileged process chains. Each agent should be assigned an owner, have a clear business purpose, a defined scope, restricted permissions, human approval for significant actions and logging that can be analysed in the security operations centre.

1. Resilient Cyber

Proactive defence measures for an organisation

The arrival of AI-driven vulnerability discovery necessitates a comprehensive overhaul of corporate security strategies. Current defensive postures, reliant on static rules and periodic human assessments, will critically fail against adversaries operating at machine speed. To navigate the new era, organisations and their CISOs must fundamentally transition their approach to resilience.

Strategic investments in cyber

The foremost proactive measure must originate in the boardroom. The era of viewing cybersecurity as a subordinate technical function, delegated to the IT department with incremental 10 percent annual budget increases, is over.[2] Chronic underinvestment is now an immediate, material business risk. To build the necessary depth of defence, organisations may need to increase their current security expenditures.

Boards must mandate the establishment of an internal AI threat management strategy. Organisations should reallocate internal experts with deep contextual knowledge of the company's bespoke network environments. This team must be equipped to point defensive AI tools at their own systems, autonomously probing for vulnerabilities ahead of adversaries.

Proactive defence measures

The CISOs must establish rigorous governance frameworks to manage the entire lifecycle of enterprise AI. Designing and implementing an organisation's AI framework based on standards, such as the NIST AI Risk Management Framework (AI RMF), ISO/IEC 27090, MITRE ATLAS and the ISO/IEC 42001 standard or regulatory guidelines, such as RBI's FREE AI Framework, is critical.

For Indian organisations, especially in the BFSI, digital payments, telecom, technology services and critical infrastructure sectors, this should be read as a board-level resilience issue. The RBI FREE-AI report sets out guiding principles and actionable recommendations for responsible AI adoption in the financial sector. These principles should be linked to cyber resilience, third-party risk, model risk, data protection and operational risk governance, rather than treated as a separate AI policy exercise.

2. The World Economic Forum

New frontier model ready architecture: Containment over patching

The Cloud Security Alliance's briefing dictates a philosophical shift for security operators. Because the window between vulnerability discovery and exploitation has closed, attempting to achieve a 100 percent patch coverage across a sprawling enterprise is an impossible endeavour. Instead, the strategic priority must pivot towards "blast radius containment." The primary objective has shifted from merely preventing initial access to structurally inhibiting the lateral movement necessary for an attacker to transform a localised, newly discovered flaw into a breach.

A resilient architecture hinges on identity as the ultimate perimeter. Threat actors use initial access to hunt for credentials, API keys and environment variables. CISOs must replace traditional long-lived credentials with short-lived, dynamically rotating tokens, effectively creating an automated "kill switch" for lateral movement.

Furthermore, legacy passwords and SMS-based multi-factor authentication are easily bypassed by AI-driven social engineering; a total transition to hardware-backed Zero Trust authentication, such as FIDO2/WebAuthn, is non-negotiable. As organisations deploy their own internal AI agents, these agents must be treated as a distinct identity class, strictly segmented and operating under absolute least-privilege permissions to prevent internal AI models from being hijacked to execute malicious commands.

Crown jewels and verified recovery

Containment needs a clear target. Organisations should maintain an operational map of crown-jewel systems, critical data stores, privileged identities, key business processes and critical third-party dependencies. This map should directly inform segmentation, privileged access, monitoring thresholds, incident response priorities and recovery sequencing. Without it, teams may know that a breach is serious but still struggle to decide what to isolate first, what to restore first and what to report under regulatory timelines.

Recovery also needs to be tested under adversarial conditions. Backups should be immutable where possible, administratively separated from the production identity plane and supported by an isolated recovery environment for the minimum viable business services. It is not enough to test whether files can be restored. Organisations should test whether they can recover safely when identity, backup administration and parts of the production infrastructure are no longer trusted.

Deploying machine-speed defences

To combat an adversary operating at machine speed, defenders must deploy equivalent technological velocity. Legacy Endpoint Detection and Response (EDR) solutions are designed to protect infrastructure, but they lack the contextual awareness to understand whether an AI model or agent is behaving maliciously.

CISOs must evaluate investing in Extended Detection and Response (XDR) and AI Detection and Response (AIDR) ecosystems. XDR federates distributed security technologies across endpoints, networks, cloud environments and identity access management. Using machine learning, XDR correlates telemetry across these historically siloed domains. For example, instantly pairing unusual endpoint login activity with anomalous network data exfiltration to trigger an automated containment response without waiting for human intervention.

AIDR represents the next necessary evolution, specifically engineered to secure the AI-native attack surface. These solutions secure the models, prompts, agents and data pipelines themselves. By utilising AIDR to monitor sanctioned and “Shadow AI” applications at runtime, organisations ensure that governance guardrails are strictly enforced from the exact moment an enterprise AI agent attempts to interact with sensitive financial systems or customer data.

Expand telemetry beyond endpoint and network signals

Machine-speed defence depends on visibility across the full attack path. Many organisations have reasonable endpoint and network telemetry but limited structured logging from custom applications, APIs, SaaS platforms, edge devices and AI agents. This creates a blind spot where AI-assisted attacks may operate. Application security logs, WAF logs, API gateway logs, SaaS audit logs, cloud control plane logs, identity events and AI-agent actions should be correlated in a shared decision layer.

Use deception to detect automated enumeration

AI-assisted attackers and autonomous tools often conduct comprehensive enumeration. That behaviour can be turned into a detection advantage through honeytokens, honey accounts, decoy credentials and carefully placed deceptive assets. Any interaction with an asset that lacks a legitimate business use can trigger a high-confidence alert. This is not a replacement for prevention, but it can give defenders an early signal when automated reconnaissance or credential harvesting begins.

Key actions and priorities

The first three months in this emerging landscape of AI accelerated cyber risk are about restoring control, not pursuing perfection. CISOs must first stabilise the environment, understand where weaknesses truly lie and build sufficient organisational capacity to address long-term resilience. The focus is clarity, containment and speed.

CISO priorities

Short term: Establish stability and visibility

Start with the board
Executive leadership needs a direct briefing on the organisation's exposure and the implications of AI driven threats. This focuses on strategic risks rather than technical details.

Form a rapid action task group
Create a compact, empowered team drawn from security, identity, cloud and legal. Their job is to triage, assess and act without bureaucratic delays.

Identify identity based weaknesses
Catalogue privileged accounts, API keys, service identities, long lived credentials and dormant access paths. These are the first routes an AI enabled attacker will target.

Check critical SaaS exposure
Confirm that high-value SaaS platforms have strong MFA, administrative logging, controlled OAuth grants and clear ownership for incident response.

Map AI touchpoints
Understand where AI systems are being used across the organisation, both officially and through shadow experimentation. Every integration point introduces data, identity and workflow risk.

Identify AI-agent permissions
For each sanctioned agent, document the systems it can reach, the operations it can perform, the identity it uses and whether human approval is required for sensitive actions.

Pause changes in sensitive environments
Legacy operational systems and critical infrastructure should not be modified until containment and rollback strategies are in place.

Medium term: Limit exposure and restrict movement

Eliminate long lived secrets

Move to short lived, auto rotating tokens and centralised secrets management. This single change dramatically reduces the scale of a compromise.

Strengthen identity authentication

Mandate hardware-based MFA for privileged users. Remove authentication mechanisms that can be easily bypassed through AI enhanced social engineering.

Review service accounts and non-SSO systems

Block interactive login for service accounts, remove unnecessary standing privileges and plan migration to managed identities, wherever possible.

Tighten segmentation

Reduce pathways around critical systems to ensure that if an attacker gains entry, the intrusion remains contained and controlled.

Unify detection across domains

Consolidate endpoint, identity, network and cloud telemetry into a single monitoring and response layer with automated containment enabled.

Onboard application, API and SaaS telemetry

Require production applications and critical SaaS platforms to send security-relevant logs to the central monitoring layer.

Apply governance to internal AI systems

Treat internal AI agents and automated workflows as privileged entities with clear boundaries and least privilege controls.

Mid-long term: Accelerate to defensive machine speed

Adopt AI aware monitoring

Visibility must extend beyond infrastructure to include prompts, models, agents and data flows, enabling detection of behaviour that traditional tooling cannot see.

Automate first response actions

Early containment, session termination, access revocation and asset isolation should occur automatically. Analysts step in after the environment has been stabilised.

Conduct continuous, controlled offensive testing

Use approved AI supported testing tools to proactively identify weaknesses. The aim is to promote swift learning cycles rather than merely ticking compliance boxes.

Adopt continuous exposure management

Combine external attack surface management, configuration posture management, attack path mapping and validation so that the organisation understands which paths lead to material impact.

Reassess third party access

Evaluate vendors, contractors and integrators who often have elevated credentials. Re scope and monitor their access more tightly.

Strengthen software and SaaS supply-chain resilience

Maintain SBOM visibility for critical applications, audit OAuth grants and ask vendors whether they use AI-assisted security testing and what SLA applies to critical internet-facing vulnerabilities.

Build a funded, long term resilience plan

Convert early insights into a 12-month roadmap focused on identity, containment and automation across security operations.

Conclusion

The emergence of highly capable, cyber-focused AI systems is driving a new phase in cybersecurity, where vulnerabilities are discovered and exploited at unprecedented speed and scale. This is a technology challenge and a board-level resilience priority, particularly for sectors such as BFSI, digital payments, telecom, technology services and critical infrastructure. AI is reshaping the nature of threats and the pace at which they evolve, making it critical to integrate AI adoption with cyber resilience, third-party risk, model governance, data protection and operational risk frameworks.

Resilience begins with clarity. Organisations must define and maintain a clear view of their crown jewels, including critical systems, data, identities and dependencies, and use this to guide protection, monitoring and response. Recovery must be real, not theoretical. It should be tested under adversarial conditions, ensuring that even when core systems and identities are compromised, the organisation can quickly and safely restore essential operations.

Cybersecurity needs to be embedded into the design of systems and decision-making processes. With AI accelerating vulnerability discovery, automating exploitation and lowering the barrier to entry for attackers, defence cannot remain static. Organisations that combine strong fundamentals with adaptive, AI-aware security architectures will be better positioned to absorb disruptions, respond effectively and recover with confidence, advancing the vision of Cyber Surakshit Bharat in this new cyber frontier.

References

1. Anthropic's Mythos moment: how frontier AI is redefining cybersecurity, https://www.weforum.org/stories/2026/04/anthropic-mythos-ai-cybersecurity/
2. Anthropic's new AI model exposes fresh risks, flaws for cybersecurity, IT services, https://timesofindia.indiatimes.com/business/india-business/anthropics-new-ai-model-exposes-fresh-risks-flaws-for-cybersecurity-it-services/articleshow/130320324.cms
3. Finance leaders warn over Mythos as UK banks prepare to use powerful Anthropic AI tool, https://www.theguardian.com/technology/2026/apr/17/finance-leaders-warn-over-claude-mythos-as-uk-banks-prepare-to-use-powerful-anthropic-ai-tool
4. The "AI Vulnerability Storm": Building a "Mythos- ready" Security Program, https://labs.cloudsecurityalliance.org/wp-content/uploads/2026/04/mythosready.pdf
5. Regulators Scrutinize Anthropic's Mythos Over Banking Cyber Risks | Let's Data Science, https://letsdatascience.com/news/regulators-scrutinize-anthropics-mythos-over-banking-cyber-r-e503e511
6. Introducing Claude Opus https://www.anthropic.com/news/claude-opus-4-7
7. Anthropic investigates report of rogue access to hack-enabling ..., https://www.theguardian.com/technology/2026/apr/22/anthropic-investigates-report-of-rogue-access-to-hack-enabling-mythos-ai
8. Commentary: Anthropic's Mythos cyber scare signals the economics ..., https://www.channelnewsasia.com/commentary/anthropic-mythos-claude-ai-models-cybersecurity-risks-6063261
9. Project Glasswing: Securing critical software for the AI era \ Anthropic, https://www.anthropic.com/glasswing/
10. Project Glasswing: Securing critical software for the AI era - Anthropic, https://www.anthropic.com/glasswing
11. Claude Mythos and the AI Cybersecurity Wake-Up Call | Bain ..., https://www.bain.com/insights/claude-mythos-and-ai-cybersecurity-wake-up-call/
12. US security agency 'found' using Pentagon 'blacklisted company' Anthropic's most powerful model yet, Mythos, https://timesofindia.indiatimes.com/technology/tech-news/us-security-agency-found-using-pentagon-blacklisted-company-anthropics-most-powerful-model-yet-mythos/articleshow/130388046.cms
13. ClaudeCode - Anthropic's Mythos Model Is Being Accessed by Unauthorized Users, https://www.reddit.com/r/ClaudeCode/comments/1ss2ibt/anthropics_mythos_model_is_being_accessed_by/

14. Project Glasswing - Anthropic, https://www.anthropic.com/project/glasswing
15. Anthropic says its most powerful AI cyber model is too dangerous to release publicly — so it built Project Glasswing | VentureBeat, https://venturebeat.com/technology/anthropic-says-its-most-powerful-ai-cyber-model-is-too-dangerous-to-release
16. New Report Details How Mythos AI Makes Cybersecurity Harder, https://www.govtech.com/security/new-report-details-how-mythos-ai-makes-cybersecurity-harder
17. AI Risk Management Framework | NIST - National Institute of Standards and Technology, https://www.nist.gov/itl/ai-risk-management-framework
18. Beyond patching: Building a Mythos-ready security program - 1Password, https://1password.com/blog/beyond-patching-building-a-mythos-ready-security-program
19. What is an AI Detection and Response (AIDR)? - CrowdStrike, https://www.crowdstrike.com/en-us/cybersecurity-101/artificial-intelligence/ai-detection-and-response-aidr/
20. What Is Extended Detection and Response (XDR)? - Palo Alto Networks, https://www.paloaltonetworks.com/cyberpedia/what-is-extended-detection-response-XDR
21. Advanced AI Threat Detection with FortiXDR | Fortinet Blog, https://www.fortinet.com/blog/business-and-technology/advanced-ai-enables-fortinets-new-fortixdr-solution
22. Anthropic Claude Mythos Preview | CrowdStrike, https://www.crowdstrike.com/en-us/blog/crowdstrike-founding-member-anthropic-mythos-frontier-model-to-secure-ai/
23. Real-time cyber safeguards on Claude | Claude Help Center, https://support.claude.com/en/articles/14604842-real-time-cyber-safeguards-on-claude
24. Reserve Bank of India, Report of the Committee to develop a Framework for Responsible and Ethical Enablement of Artificial Intelligence in the Financial Sector, https://www.rbi.org.in/scripts/bs_pressreleasedisplay.aspx/BS_PressReleaseDisplay.aspx?prid=61018

Connect with us

Sathish Gopalaiah
President - Consulting
Deloitte South Asia
sathishtg@deloitte.com

Gaurav Shukla
Partner and Leader - Cyber
Deloitte South Asia
shuklagaurav@deloitte.com

Ashish Sharma
Partner
Deloitte India
sashish@deloitte.com

Anand Tiwari
Partner
Deloitte India
anandtiwari@deloitte.com

Munjal Kamdar
Partner
Deloitte India
mkamdar@deloitte.com

Mayuran Palanisamy
Partner
Deloitte India
mayuranp@deloitte.com

Contributors

Sanjeev Singh

Hiten Panchal

Sunita Kumari

Deloitte.