Deloitte.

Together makes progress

Machine-Speed Cyber Risk

What EMEA Financial Services leaders need to know

May 2026

Machine-Speed Cyber Risk

What Financial Services leaders need to know

A step-change in artificial intelligence capability may significantly compress the time between identifying a cyber vulnerability[1] and exploiting[2] it. Early evidence from the AI company Anthropic - including its unreleased Claude Mythos model and the restricted-access Project Glasswing[3] initiative - suggests that AI systems may now be capable of identifying vulnerabilities across major software platforms at scale. While these claims remain only partially validated,[4] the implication is clear: it has never been so important to get the basics right and execute existing cyber efforts at pace and scale.

Anthropic and Project Glasswing

Anthropic claims that its latest Claude Mythos model has demonstrated the ability to autonomously identify cyber vulnerabilities across major operating systems and web browsers. As a result, Anthropic has not yet publicly released its latest model, but in early April 2026 announced Project Glasswing, an invitation-only defensive cyber security programme. The project was initially limited to a small number of launch partners, and has since been extended to a reported 40 additional organisations that build or maintain critical software.

By giving selected companies' security teams access to Mythos before attackers acquire comparable capability, Anthropic creates a window for cyber defenders to find and remediate weaknesses in their own systems.

The strongest public claims of capability still originate from Anthropic itself, with public proof trailing the private disclosure process. At this stage, purported capabilities should be treated as provisional, and financial institutions should avoid treating every headline-grabbing claim as verified fact.

Project Glasswing should, nonetheless, be taken seriously. While the rate at which vulnerabilities are being discovered has been increasing for many years,[5] there is enough public information to treat this as a material capability jump. Other AI and tech companies are likely to be developing similar capabilities, even if public evidence varies by provider.

Regardless of the Anthropic specifics, the direction of travel is clear: AI capabilities in cyber will continue to advance rapidly. Future models are likely to exceed the capabilities Claude Mythos is purported to have today and will likely be widely available, which fundamentally impacts cyber threats.

These AI capabilities will compress the time between vulnerability discovery, disclosure,[6] and exploitation, which raises implications around operational resilience,[7] remediation prioritisation, and governance. In some cases, working exploits may exist before disclosure[8] and also be more widely disseminated than historical norms. Boards should treat this as a time compression risk that may threaten the resilience of critical services, rather than a narrow technology story.

How cyber threats might change

The constraint on sophisticated cyber attacks is shifting from expertise to access. The most advanced cyber exploit development capabilities have historically been concentrated in a small set of highly capable state and criminal actors. As AI models improve, the capability for advanced cyber attacks will depend less on human skill and more on money, access, data, and the willingness to operationalise. A wider range of attackers are likely to gain access to sophisticated cyber techniques.

This is a dual-use development. Restricted access to cyber-capable models may give some defenders a temporary advantage, but it also creates clear geopolitical and security risks. Advanced capability is unlikely to remain confined to defenders; over time it may be accessed, replicated, or acquired by malicious attackers, increasing both the scale and sophistication of attacks.

From financial services organisations' point of view, the same capability jump that elevates cyber threats may ultimately strengthen cyber defence. A step change in AI models' ability to find vulnerabilities can ultimately be good for cyber security, because firms and their software suppliers can use AI to identify and fix cyber weaknesses throughout product lifecycles. While this ultimately strengthens defensive capability, it does create a near-term asymmetry where vulnerability discovery is likely to outpace remediation for a period.

Controlling access to these technologies is inherently difficult. These models can be developed across multiple jurisdictions, subject to differing legal frameworks, export controls, and regulatory priorities. There is no single global authority able to enforce consistent rules, and coordination between governments is likely slower than the pace of technological change. Once capabilities begin to diffuse - through open-source release, model leakage, or commercial competition – attempts at restrictions will become progressively less effective.

The debate has therefore already shifted from *"Is this model dangerous?"* to *"Who controls access to this capability, and for how long?"* As this evolves, firms should assume that advanced capabilities will become more widely available and plan on that basis.

Why EMEA financial services is exposed

Financial services organisations may be disproportionately exposed to these capability leaps because the industry combines high-value targets, large, complex legacy technology estates, common vendors, significant third-party reliance and connectivity, strict uptime expectations, and heavy regulation.

The strategic issue is not that AI creates magical new bugs. It exposes existing vulnerabilities faster than firms can remediate them. This creates not only firm-level risk, but also potential systemic risk where common vulnerabilities are exploited rapidly across institutions sharing infrastructure or suppliers. If financial services organisations aren't already focusing on improving basic cyber capabilities, particularly in vulnerability management,[9] Project Glasswing should serve as a clear signal of urgency.

The practical consequence for firms is not a radical change, but a significantly increased importance of getting the basics right and executing existing efforts at pace.

The regulatory backdrop in EMEA points in the same direction. In the EU, the European Banking Authority has aligned its ICT guidance around the Digital Operational Resilience Act, and ECB Banking Supervision has made cyber security, third-party risk management, threat-led penetration testing and ICT change management explicit supervisory priorities for 2026-28.[10] The UK's financial services regulators require firms to operate important business services[11] within impact tolerances.[12] These regulations were not designed for AI-enabled vulnerability discovery, but become more relevant in a world where vulnerability discovery speeds up sharply.

For financial services, the practical message is not *"new regulation is coming for Mythos/Glasswing."* It is *"your existing cyber hygiene, operational resilience, third-party risk and cyber expectations just became less forgiving."*

What does not change

While the threat model changes, the cyber control agenda broadly does not. If it was not already considered urgent, getting the cyber fundamentals right with speed and consistency is now even more critical. This distinction is critical for boards: the control framework remains familiar, but the pace and tolerance for delay do not.

That is why the institutions most at risk are not necessarily those with the fewest AI pilots or the lowest AI adoption. They are those with slow patch cycles, weak identity and privileged access control,[13] unsegmented networks, limited third-party visibility, unsupported or end-of-life technology,[14] and inadequately tested recovery plans.

The appropriate response continues to be to reduce unnecessary exposure; apply security updates rapidly, manage access, segment networks, monitor for malicious activity and respond quickly. This is not so much a radical change, but a harsher service-level agreement for the old agenda.

What does change

What does change is the premium on automation, including AI-assisted automation. Finding, prioritising, and addressing vulnerabilities needs to become continuous. Financial services organisations need to move faster on the basics of vulnerability management, supported by AI tools, deployed in a controlled way.

While the consequences of AI-enabled cyber defence will be felt throughout the security operation, they are likely most immediately visible in vulnerability management.

Automation without control introduces a new class of failure: rapid, systemic misconfiguration at scale. Safeguarding and tightly controlling AI deployments used for cyber defence is critical, and regulators such as FINMA have started publicly warning of the potential risks with uncontrolled deployments.[15] AI-driven tools may prioritise patches, automate remediation, or even trigger changes across large estates; without robust governance, this creates a risk of propagating errors at scale, introducing misconfigurations, or amplifying adversarial manipulation. Effective controls such as human-in-the-loop[16] approvals for high-impact changes, strong validation of training data and model outputs, access management, and continuous monitoring need careful attention to ensure that AI augments rather than undermines cyber efforts.

Short-term vulnerability management considerations

In the short term, financial services organisations should consider a vulnerability patching sprint. In practice that means building an up-to-date heatmap of end-of-life assets, externally exposed systems, crown-jewel systems or data stores, and important business services. Such a sprint should measure actual remediation latency, not just policy targets. It should also identify where approvals, testing windows, ownership ambiguity or vendor dependence cause delay.

This sprint is only a temporary exposure-reduction measure. Tighter patch SLAs, more automation in the scanning pipeline, and larger maintenance windows are the organisational equivalent of running faster on a treadmill. They buy down immediate exposure to a point, but largely address the symptom rather than the cause. In our view, no realistic permanent increase in team size or operational efficiency can close the vulnerability gap efficiently and sustainably.

If not already in progress, security and engineering teams should start using current AI models in controlled ways to detect vulnerabilities now, rather than wait for broader access to Mythos-class systems. The immediate objective is not to put AI at the centre of the security operating model. It is to use AI pragmatically to find and remove the easily exploitable weaknesses an AI-assisted attacker would otherwise exploit first.

Medium-term vulnerability management considerations

In the medium term, the backlog is likely to grow faster than it can be cleared, at least for a period. The short-term sprint is only a temporary exposure-reduction measure. Financial services organisations need to move away from reactive and episodic patch and vulnerability management, all too common today, towards continuous vulnerability identification, triage and remediation, supported by appropriate AI tools and agents.

Critically, this shift must be accompanied by a more intelligent approach to prioritisation. The challenge is no longer identifying vulnerabilities - it is deciding which ones matter and acting quickly enough. Firms should move beyond generic prioritisation scores to a more contextual, threat-informed model that considers aspects such as the business criticality of services, how adversaries actually operate, and the effectiveness of existing controls - effort needs to be focused where it most reduces material risk. This is a shift from measuring success by patch speed to measuring success by risk reduction, and reflects how attackers behave in practice—targeting exposed, high-value weaknesses rather than the highest theoretical severity scores.

Over time, firms may need narrowly scoped autonomous tools that continuously test systems, update risk scores, and escalate exceptions to human reviewers within clearly defined guardrails.[17] These agents do not replace human judgement on significant decisions; they operate within defined boundaries and must be secured appropriately.

Board implications

In financial services organisations, boards typically carry explicit accountability for cyber risk as a core component of prudential soundness and operational resilience, with regulators increasingly expecting or requiring it to be a standing board-level agenda item. Expectations from the ECB, FCA, PRA, FINMA and regional supervisors in the Middle East are converging on the need for demonstrable board oversight.

Boards should treat artificial intelligence's likely impact on cyber risk as an operational-resilience issue with cyber and AI dimensions, not as a niche technical topic. The right questions are evolving, but should include:

- Which critical services currently carry the highest level of cyber risk and why?
- Of the vulnerabilities we know about, how many affect critical or externally exposed services, and how serious are they?
- How quickly do we move from knowing about a vulnerability, to making a risk-based decision, putting interim protections in place, and fully fixing it – and how has that changed in the past 12 months?
- Where do we have outdated technology, weak access controls, or insufficient visibility into the components and dependencies underpinning our critical services?
- What are our controls over developer use of external and AI-generated code?
- Which suppliers are most critical to our services, and how confident are we in their ability to manage vulnerabilities and keep systems supported?
- Where is AI is used to influence cyber-related decisions, and how do we ensure appropriate human oversight?
- What evidence do we have that our prioritisation model reduces real-world risk rather than simply improving patching metrics?
- If a highly capable attacker had access to the same AI tools we are piloting, where would they successfully compromise our business today?
- Where a vulnerability affects an important business service, can we remain within our agreed impact tolerances while it is being remediated?

Longer term outlook

The longer-term outlook is more positive than the headlines may suggest. Defenders may finally have a chance to win decisively, but only if financial services organisations use this window to shorten remediation cycles before these capabilities become widely available.

The firms that will cope best will not be those that simply try to patch everything faster. They will be the firms that know which services matter most, which vulnerabilities are actually exploitable in their environment, where third-party and open-source dependencies sit, and where AI can safely accelerate detection and remediation within clearly defined limits. The board's role is to ensure management can show that context, act on it quickly, and escalate exceptions before they threaten critical-service resilience thresholds, prudential soundness, or customer outcomes.

Authors

Nick Seaver
United Kingdom

João Luis Fonseca
Portugal

Diego Giordano
Italy

Volker Burgers
Germany

Delisa Stone
Netherlands

Wil Rockall
United Kingdom

Local contact – Hungary

Zoltán Szöllősi
Partner
Cyber Strategy & Transformation Leader
Deloitte CE
zszollosi@deloittece.com

László Tóth
Partner
Cyber Defense and Resilience Leader
Deloitte CE
ltoth@deloittece.com

Endnotes

1 A weakness in systems, software, or processes that can be used by an attacker to cause harm, such as disruption, data loss, or unauthorised access.
2 Taking advantage of a vulnerability in practice.
3 https://www.anthropic.com/glasswing
4 https://cetas.turing.ac.uk/publications/claude-mythos-future-cybersecurity
5 https://www.nist.gov/news-events/news/2026/04/nist-updates-nvd-operations-address-record-cve-growth
6 The point at which a vulnerability becomes known to the organisation, supplier, or publicly.
7 The ability of a firm to continue delivering its most important services through disruption, within acceptable limits.
8 Commonly called a "zero-day exploit" as the security flaw is both unknown to the vendor or financial services organisation, and has not yet been patched, leaving "zero days" to fix it before it can be exploited.
9 The process of identifying, assessing, prioritising, and fixing security weaknesses across systems.
10 https://www.bankingsupervision.europa.eu/framework/priorities/html/ssm.supervisory_priorities202511.en.html
11 The services a firm provides where disruption could cause significant harm to customers, markets, or the firm. Defined under frameworks such as the UK's PRA operational resilience regime.
12 The maximum level of disruption a firm is willing to tolerate for an important business service.
13 Controls that restrict and monitor powerful system access rights that could significantly impact systems or data if misused.
14 Systems or software that are no longer supported by the supplier and do not receive security updates.
15 https://www.finma.ch/en/~/media/finma/dokumente/dokumentencenter/myfinma/4dokumentation/finma-aufsichtsmitteilungen/20241218-finma-aufsichtsmitteilung-08-2024.pdf
16 A control where a person must review and approve certain actions before they are carried out by automated or AI systems.
17 Guardrails are controls that limit what AI systems can do, ensuring they operate safely, predictably, and within defined boundaries.
18 Software code produced with the assistance of artificial intelligence tools rather than written entirely by developers.

Deloitte.

Together makes progress