BIS

BIS Bulletin

No 129

A Mythos moment? Frontier AI and cyber risk

Iñaki Aldasoro, Raphael Auer, Jon Frost, Fernando Pérez-Cruz

July 2026

BIS Bulletins are written by staff members of the Bank for International Settlements, and from time to time by other economists, and are published by the Bank. The papers are on subjects of topical interest and are technical in character. The views expressed in this publication are those of the authors and do not necessarily reflect the views of the BIS or its member central banks.
The authors thank Pablo Hernández de Cos, Byeungchun Kwon, Ulf Lewrick, Randy Miskanic, Vasily Pozdyshev, Kumar Rishabh, Vatsala Shreeti, Andras Valko and Leanne Zhang for helpful comments and suggestions, and Rudraksh Kansal, Byeungchun Kwon and Vivekananda Allam for excellent statistical assistance. They thank Nicola Faessler for administrative support.

The editors of the BIS Bulletin series are Gaston Gelos and Frank Smets.

This publication is available on the BIS website (www.bis.org).

ISSN: 2708-0420 (online)
ISBN: 978-92-9259-970-6 (online)

Iñaki Aldasoro
inaki.aldasoro@bis.org

Raphael Auer
raphael.auer@bis.org

Jon Frost
jon.frost@bis.org

Fernando Pérez-Cruz
fernando.perezcruz@bis.org

A Mythos moment? Frontier AI and cyber risk

Key takeaways

- *Frontier artificial intelligence (AI) models increase the speed, scale and complexity of cyber attacks, and they also strengthen cyber defence. But the costs are asymmetric and may favour attackers.*
- *The medium-term impact on systemic cyber risk depends on the various actors' access to the most advanced tools, on compute power to run the tools and on economic incentives.*
- *Given the pace of recent developments, swift adoption of frontier AI models to review code bases and fix vulnerabilities is essential. International coordination can support authorities in addressing these issues.*

Among the most notable capabilities of frontier artificial intelligence (AI) models is their capacity to find vulnerabilities in software and hardware systems. Anthropic's Mythos, announced on 7 April 2026 and released only to select partners, is a case in point (Carlini et al (2026)). Mythos – and soon thereafter other frontier AI models, such as OpenAI's GPT-5.5 – is able to not only identify cyber vulnerabilities but also develop exploits to take advantage of them, and to autonomously carry out sophisticated multi-step, multi-vulnerability cyber attacks. Does this represent a "Mythos moment" – a wake-up call that requires a fundamental reconsideration of views on the robustness and resilience of financial market infrastructures?

The financial system is an obvious place of concern for information technology vulnerabilities. Banks, payment systems and other market infrastructures are among the most heavily targeted and most densely interconnected parts of the economy. They depend on long, complex chains of proprietary and open source software as well as third-party suppliers. A step change in attackers' capabilities could therefore have consequences that extend well beyond any single institution and bear directly on financial stability.

This Bulletin sets out potential channels through which frontier AI models can affect cyber security risk at scale. It discusses the impact of new tools on the capabilities and incentives of both attackers and defenders and draws on available data. Finally, it discusses potential public policy responses to support financial stability.

Performance of frontier AI models on cyber security tasks

Frontier AI models have significantly advanced cyber offensive capabilities. For instance, recent evaluations by the UK government's AI Security Institute (AISI) suggest that frontier models are improving cyber-relevant offensive tasks. AISI tests models on a suite of 95 narrow cyber tasks across four difficulty tiers in a "capture-the-flag" benchmark setting, in which a model must find and exploit a deliberately planted vulnerability to retrieve a hidden token. Mythos achieved a 68.6% pass rate for expert-level tasks – higher than any model before (AISI (2026a); Folkerts et al (2026)). Yet these capabilities do not appear to be unique to Mythos. OpenAI's GPT-5.5, released just weeks after Mythos, performed even better, with a 71.4% pass rate (Graph 1.A; AISI (2026b)). Moreover, in a cyber range, ie long multi-step attack simulations, both Mythos and GPT-5.5 were able to achieve a full network takeover in some attempts (Graph 1.B).

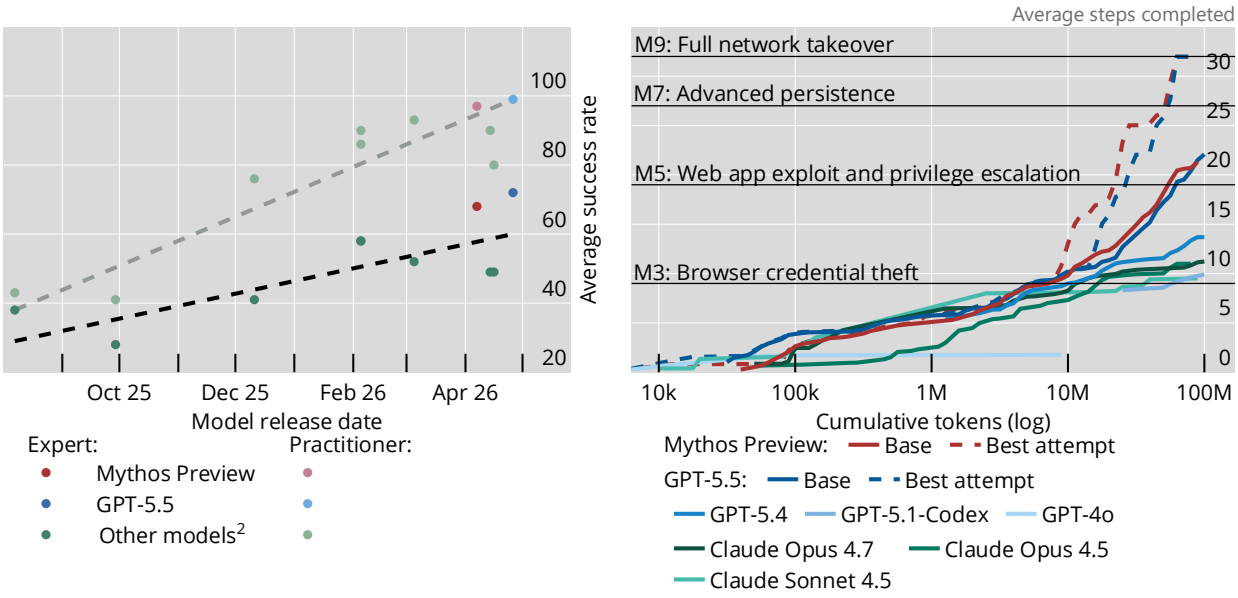
Cyber offence is unusually well suited for frontier AI tools. Unlike many tasks in which an AI system must navigate open-ended ambiguity, an attack unfolds as a structured, sequential workflow, codified in widely used industry frameworks (Hutchins et al (2011); Strom et al (2020)). Each stage generates machine-readable outputs (system logs, error messages, code, catalogued vulnerabilities) that a model can parse, reason over and act upon, with the success or failure of one step furnishing immediate feedback for the next. These tight feedback loops, comparatively rare in less structured domains, are precisely the conditions under which such models learn and improve most rapidly. The inflection point, which Mythos was the first model to reach, arrives not when a model can identify a single exploit, but when it can reliably link steps and adapt them to targets with limited human oversight. Encouragingly, the same properties operate in reverse: defenders can apply identical reasoning to anticipate, detect and remediate intrusions, so advances in model capabilities need not accrue to attackers alone.

Frontier AI models show impressive cyber capabilities

Graph 1

A. Mythos and GPT-5.5 show superior performance on expert- and practitioner-level cyber tasks[1]

B. Mythos was the first model to complete all steps in a cyber range, and GPT-5.5 showed similar performance[3]



[1] Measured pass rate (y-axis) of selected models on an advanced capture-the-flag challenge with a 50 million token budget, by date (x-axis). [2] GPT-5, Claude Sonnet 4.5, Codex 5.2, Claude Opus 4.6, Codex 5.3, GPT-5.4, GPT-5.4 Cyber, Opus-4.7. Dashed lines are a least-square fit for the expert and practitioner tasks, respectively [3] Mean number of steps completed in the 32-step corporate network attack simulation "The Last Ones", over 10 attempts, with a 100 million token budget per attempt. "M3", "M5", etc refer to selected milestones in the attack chain. The dashed lines represent the outcome of the best attempt (ie a full network takeover succeeded).

Source: AI Security Institute (2026b).

There are costs to mounting such an attack, but they are not prohibitive. A full attack chain consumes approximately 100 million tokens (a unit of input or output data). At current cloud prices, running such an attack on Claude Mythos Preview would cost some $5,000–$10,000, depending on the balance between input tokens (the code and data supplied to the model) and the more expensive output and reasoning tokens used to plan and execute it. Other frontier models such as Claude Opus 4.8, Gemini 3.5 and GPT 5.5 would cost about a fifth as much, and cheaper models such as Grok or DeepSeek can cost as little as $50–$100 per attack. As these costs fall, it becomes easier for less sophisticated actors to mount attacks.

These developments have renewed public concern and policy discussions around cyber risk. Supervisors and banks have aired these concerns publicly: some central banks are reported to have questioned banks about their exposure to new models, and senior bank executives have described them as a serious threat while warning that more threats will follow (see eg Canepa (2026)).

Incentives of attackers and defenders

Whether frontier AI tools raise or lower systemic cyber risk depends on who gains access to them and the economic incentives they face. Cyber attackers are a heterogeneous group. State and state-sponsored actors pursue espionage, disruption (eg cyber warfare) or strategic advantage. Organised criminal groups seek financial gain, eg through ransomware and fraud. Firms may engage in industrial espionage. And individuals act for profit, notoriety or ideology. Alongside this sits a black market in which zero-day vulnerabilities (ie hidden flaws or bugs in a system that are unknown to the developer), stolen credentials and ready-made hacking tools are traded. At the same time, a sizeable and entirely legal market exists for cyber security and vulnerability discovery. In this latter market, firms seeking to defend themselves run bug bounty programmes that pay researchers to find and disclose flaws so that they can be fixed.[1]

Frontier AI models are being adopted by both attackers and defenders. These models can lower the cost of finding and exploiting vulnerabilities for attackers but can also help defenders discover and fix them.[2] The economic costs are, nonetheless, asymmetric: a defender must protect every system continuously, whereas an attacker needs only one viable route in. Tools that cut the cost to find and exploit that route can therefore shift the balance towards offence even when both sides adopt frontier AI models at the same pace. Market intelligence also suggests that cyber defence strategies may increasingly imply larger costs for compute, given the need to detect and respond to (cheaper and more frequent) attacks.

The cyber defence response is already visible in the data. The release of ever more capable AI tools has coincided with a marked increase in security bug fixes. For example, monthly bug fixes by Firefox jumped sixfold after the announcement of Mythos Preview (Graph 2.A). It is an open question whether the pace of fixes will remain structurally higher or will fall again as bugs are found and remedied. This will depend on the capacity of new models to find new bugs or alternatives for attacks.

More generally, the pace at which new vulnerabilities are discovered and catalogued has been rising steadily for years and is now accelerating. This fits a longer-run pattern in the development of cyber tools. The number of critical common vulnerabilities and exposures (CVEs) published has been rising each year (Graph 2.B).[3] A CVE is an identifier for each vulnerability discovered in a software system; each is assigned a Common Vulnerability Scoring System (CVSS) score that indicates its severity. A score climbs toward 10 when a flaw is easy to exploit, requires no special access and would cause severe damage, such as allowing full remote control of a machine. CVEs with a CVSS score of nine or higher are deemed critical. The number of critical CVEs reported has risen from about seven per day over 2018–21 to 10 per day in 2022–25, to almost 20 per day after 1 April 2026.

Broadly speaking, data on CVEs fall into three distinct periods, each coinciding with a wave of AI-assisted coding tools. Between the first two periods (2018–21, 2022–25), the discrete jump in the number of CVEs may relate to the availability of GitHub Copilot (released in June 2021) or OpenAI's GPT application programming interface (API) (first available in June 2020).[4] Over 2022–25, the annual count was broadly stable, as developers became accustomed to coding-assistant tools. The jump in reported CVEs in the second quarter of 2026 may relate to Mythos and other models for coding becoming available. But these counts should be interpreted with care: the number of published CVEs also reflects the growing size and reach of the software industry and the size of the security research community, not AI capabilities alone.

[1] Another consequence of the availability of frontier AI models is that most bug bounty programmes are reducing payouts and are being reconsidered overall. See M Jackson, "Bug bounty isn't dead, but the old model is breaking", *Aikido*, 13 April 2026, www.aikido.dev/blog/bug-bounty-isnt-dead.

[2] Aldasoro et al (2024, 2025) discuss how generative AI can strengthen both cyber defence and offence, especially in social engineering, monitoring and data security settings. The same is likely to hold for frontier models (Carlini et al (2026)). Some argue that the capabilities and applications of AI in attacks exceed those for defence (Potter et al (2025)).

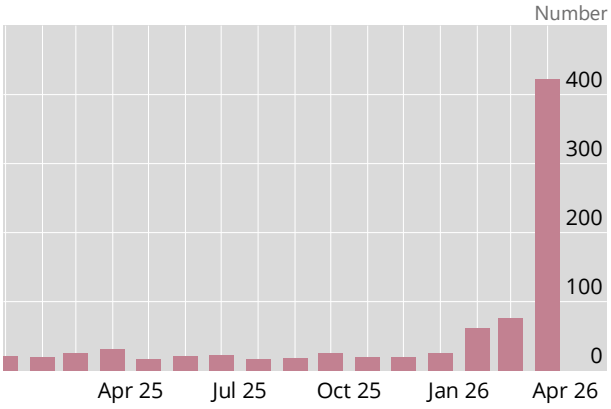
[3] The data for the CVE were downloaded from the National Institute of Standards and Technology (NIST) National Vulnerability Database (NVD) REST API v2 (available at https://services.nvd.nist.gov/rest/json/cves/2.0, accessed on 18 May 2026).

[4] Unlike ChatGPT's release in November 2022, the API access to GPT went mostly unnoticed.

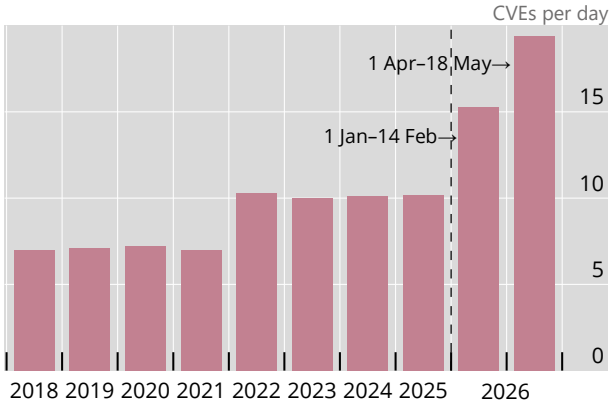
Security bug fixes and capabilities to identify vulnerabilities are growing

Graph 2

A. Firefox security bug fixes saw a notable jump after the announcement of Mythos Preview [1]



B. Mean of daily critical CVEs (CVSS >= 9.0) reported was stable in 2022–25, but jumped up in 2026[2]



CVEs = common vulnerabilities and exposures

[1] Firefox security bug fixes by month (all sources, all severities). [2] Mean of daily critical CVEs (ie those with a Common Vulnerability Scoring System (CVSS) score >= 9.0) per calendar year.

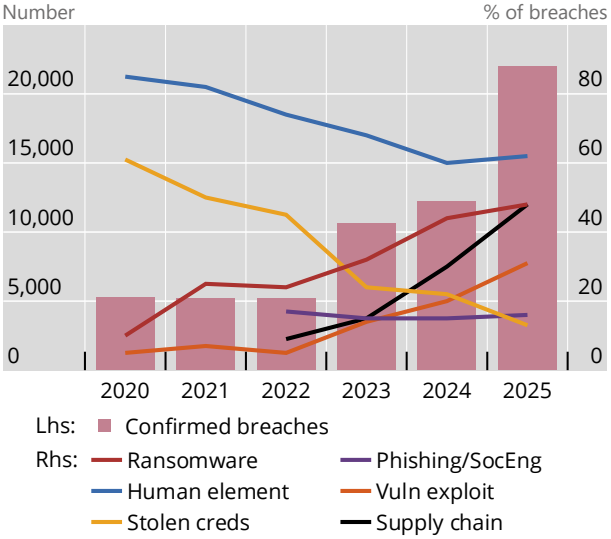
Sources: NIST National Vulnerability Database API (accessed 18 May 2026), BIS calculations.

Reported data breaches have risen in step and a growing share are fully automated. By a global estimate, 2025 recorded a significant jump (Graph 3.A), and breaches attributed to automated rather than human action became more common. These data are incomplete, however, and tend to reveal more about the type of incident than about the specific tools used, not least because attackers do not self-report.

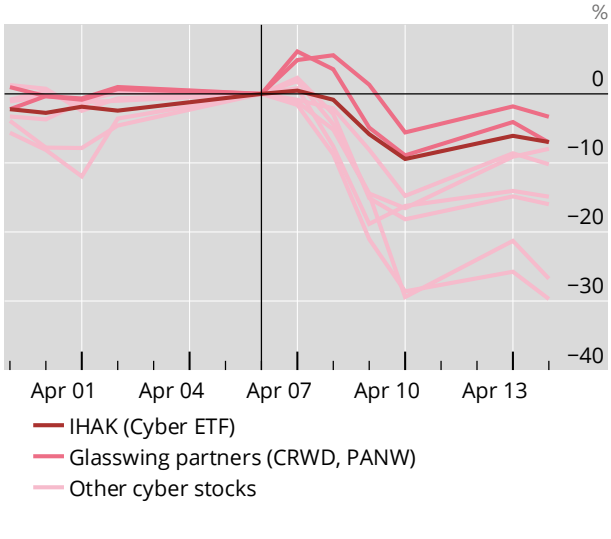
Data breaches have been rising; since Mythos, market expectations have shifted

Graph 3

A. The number of reported data breaches has increased over time, and fewer have a human element[1]



B. Investor views have shifted, and cyber-related stocks saw negative cumulative abnormal returns [2]



[1] Based on global data breaches collected from domestic and international law enforcement, forensic firms, law firms, cyber insurers, cyber security industry sharing groups and the Verizon Threat Research Advisory Center caseload. [2] Cumulative abnormal returns (CARs) are computed from a market model regression of each stock's daily log returns on Nasdaq 100 log returns, estimated over 2 December 2025 to 20 March 2026 (ie preceding the initial leak of information about Mythos). Abnormal returns are the residuals from this model evaluated over the event window; CARs are daily and are normalised to zero on 6 April 2026, so values from 7 April onwards measure cumulative abnormal performance since the eve of the announcement. "Glasswing partners" refers to selected known companies that were granted access to Mythos (Crowd Strike (CRWD) and Palo Alto (PANW)).

Sources: Verizon; LSEG Datastream.

Market reactions provide a complementary, forward-looking signal. The cumulative abnormal stock returns of a number of cyber security firms fell sharply with the announcement of Mythos (Graph 3.B). This was widely interpreted as a reflection of investor concerns about obsolescence of existing cyber security products and services.[5] This is a notable reaction, given that a more dangerous threat environment might instead be expected to raise demand for such products.

Potential public policy responses

The rapid evolution of AI-driven cyber attack capabilities calls for coordinated public policy responses. The private return to finding and disclosing a vulnerability may be smaller than its system-wide benefit, especially when many institutions share the same code or supplier. Firms may therefore underinvest in discovery and disclosure. This externality provides a direct rationale for coordinated scanning and information-sharing. In both the short and medium term, domestic and international coordination can mitigate risks to critical infrastructure and sustain trust in the stability of the financial system.

An initial safeguard is domestic cooperation, not least to ensure that defenders gain access to frontier models before attackers. The financial industry, central banks and financial supervisors can collaborate with national security agencies and other stakeholders to accelerate vulnerability remediation, strengthen core cyber hygiene practices and address exposures from in-house software dependencies, legacy systems, open source code, third parties and internet-facing assets. The same tools that lower costs for attackers can be used defensively, to find and fix flaws in firms' own systems before others do. Engaging critical suppliers, including cloud providers, and updating exercises to simulate third-party software compromises and compressed attacker timelines are also important.

Given the global nature of the risks involved and the interconnected nature of the global financial system, cross-border information-sharing is indispensable. Of course, there can be geopolitical constraints to cooperation in some areas, eg threat intelligence. Still, global forums can build on existing operational resilience frameworks, such as the Principles for Operational Resilience and Effective Practices for Cyber Incident Response and Recovery (FSB (2020, 2021)), or discuss AI-related operational risks reflected in risk identification, scenario analysis and resilience planning (BCBS (2021)). For financial market infrastructures, AI-driven threats can be mapped across the pillars to identify, protect, detect, respond and recover in the CPMI–IOSCO cyber resilience guidance (CPMI–IOSCO (2016)). The G7 (2026) already made a commitment to enhance information-sharing and identify best practices. It has also undertaken to reinforce the preparedness of the financial sector and to map the cyber security risks related to AI models. The cyber security agencies of the Five Eyes (2026) have also urged leaders to reduce attack surfaces and accelerate patching processes.

In the medium term, resilience frameworks must adapt to the evolving threat environment shaped by AI. Regulators and standard-setting bodies can translate existing principles into practical toolkits for extreme scenario design, such as AI-enabled attack paths, and update them as threats evolve. Strengthened oversight of third-party and supply-chain risk is increasingly important as digital dependencies deepen. Operators of systemic infrastructures could be encouraged to participate in pretesting of new AI models, while volunteer pilots and shared assessments can accelerate collective learning. Demand for capacity building is likely to rise;[6] international initiatives can prototype AI-enhanced resilience tools and support jurisdictions with varying resources. Information-sharing arrangements may need recalibration to improve timeliness, reciprocity and legal protections, supported by interoperable taxonomies and secure channels.

[5] An alternative, more benign interpretation is possible: demand for cyber security firm products may fall because the new AI models strengthen in-house cyber defences.

[6] At the same, capacity building may need to recognise a complementarity between access to AI tools and in-house organisational capital. Implementation alone does not insulate firms' innovation from cyber risk; resilience against such risk is strongest when implementation is paired with internal AI development (Rishabh et al (2026)).

The stakes for the financial system are especially high. Payment, trading and post-trade platforms operate at a very large scale and are tightly interconnected with billions of firms and individuals. Complex software stacks and intricate supply chains complicate patching and monitoring, and disruptions can be hard to detect and unwind. These same infrastructures are an attractive target for actors seeking to inflict large-scale economic damage. By learning, adapting and communicating effectively, authorities and the financial industry can turn this pivotal moment into durable gains in global cyber resilience.

References

AI Security Institute (AISI) (2026a): "Our evaluation of Claude Mythos Preview's cyber capabilities," 13 April.

——— (2026b): "Our evaluation of OpenAI's GPT-5.5 cyber capabilities", 30 April.

Aldasoro, I, S Doerr, L Gambacorta, S Notra, T Oliviero and D Whyte (2024): "Generative artificial intelligence and cyber security in central banking", *BIS Papers*, no 145.

Aldasoro, I, L Gambacorta, A Korinek, V Shreeti and M Stein (2025): "Intelligent financial system: how AI is transforming finance", *Journal of Financial Stability*, vol 81, 101472.

Basel Committee on Banking Supervision (BCBS) (2021): *Principles for operational resilience*.

Canepa, F (2026): "ECB to quiz bankers about risks of Anthropic's new AI model, source says", *Reuters*, 15 April.

Carlini, N, N Cheng, K Lucas, M Moore, M Nasr, V Prabhushankar, W Xiao et al (2026): "Assessing Claude Mythos Preview's cybersecurity capabilities", Anthropic, 7 April.

Committee on Payments and Market Infrastructures–International Organization of Securities Commissions (CPMI–IOSCO) (2016): "Guidance on cyber resilience for financial market infrastructures", *CPMI Papers*, no 146.

Financial Stability Board (FSB) (2020): *Effective practices for cyber incident response and recovery*.

——— (2021): "Principles for operational resilience", *Compendium of Standards*.

Five Eyes (2026): "Five Eyes cyber security agencies statement", 22 June.

Folkerts, L, W Payne, S Inman, P Giavridis, J Skinner, S Deverett, J Aung, E Zorer, M Schmatz, M Ghanem, J Wilkinson, A Steer, V Hong and J Wang (2026): "Measuring AI agents' progress on multi-step cyber attack scenarios", *arXiv*, 10.48550/arXiv.2603.11214.

G7 (2026): "G7 Finance Ministers' and Central Bank Governors' Communiqué", 19 May.

Hutchins, E, M Cloppert and R Amin (2011): "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains", in J Ryan (ed), *Leading issues in information warfare and security research*, vol 1.Potter, Y, W Guo, Z Wang, T Shi, H Li, A Zhang, P G Kelley, K Thomas and D Song (2025): "Frontier AI's impact on the cybersecurity landscape", *arXiv*, 10.48550/arXiv.2504.05408.

Rishabh, K, R Mihet and J Jang-Jaccard (2026): "Cyberrisk and AI Firms", *Review of Corporate Finance Studies,* cfag018.

Strom, B, A Applebaum, D Miller, K Nickels, A Pennington and C Thomas (2020): "MITRE ATT&CK: design and philosophy", MITRE Corporation technical report, no MP180360R1.