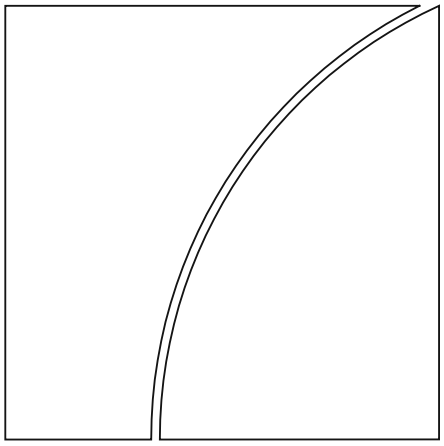
Financial Stability Institute | BIS



Occasional Paper

No 28

When machines attack: frontier AI cyber threats and policy responses in the financial sector

by Juan Carlos Crisanto, Adrien Currat and Jeffery Yong

September 2026

JEL classification: G21, G22, G28

Keywords: artificial intelligence, frontier AI, cyber security, cyber risk, operational resilience

FSI Occasional Papers aim to contribute to international discussions on a wide range of topics of relevance to the financial industry and its regulation and supervision. The views expressed in this publication are those of the authors and do not necessarily reflect the views of the BIS, its member central banks or the Basel-based standard-setting bodies.

Authorised by the Chair of the FSI, Fernando Restoy.

This publication is available on the BIS website (www.bis.org). To contact the BIS Global Media and Public Relations team, please email media@bis.org. You can sign up for email alerts at www.bis.org/emailalerts.htm.

© *Bank for International Settlements* 2026. *All rights reserved. Brief excerpts may be reproduced or translated provided the source is stated.*

ISSN 1020-9999 (online)

Abstract

Frontier artificial intelligence (AI) models are a game changer in the cyber threat landscape. Unlike earlier generations of AI models, they can autonomously identify critical vulnerabilities, develop effective exploits and conduct increasingly complex multi-step cyber operations, thus reducing the expertise, time and resources needed to carry out sophisticated attacks. At the same time, these capabilities offer significant defensive opportunities, including faster vulnerability discovery, threat detection and incident response.

The risks for financial institutions arise from compressed cyber remediation windows, higher likelihood of breach and amplified third-party dependencies. By collapsing the window from discovery to exploitation and automating exploit chaining, frontier models materially increase the likelihood of breach, with unpatched software becoming the leading initial access vector in many incidents. Reliance on common cloud, software and frontier AI providers introduces concentration and sovereign access risks, whereby a single provider's disruption or policy decision can cascade across firms and jurisdictions.

Financial authorities are converging on a pragmatic response. Rather than introducing new AI-specific cyber regimes, they are reinforcing existing cyber risk management and operational resilience frameworks while adapting supervisory expectations to the new cyber threat environment. Policy responses increasingly emphasise governance capable of supporting timely decision-making, accelerated patching and enhanced response and recovery capabilities. Frontier AI therefore does not fundamentally change the foundations of cyber resilience, but it significantly increases the speed and intensity with which established practices need to be executed.

Contents

When machines attack: frontier AI cyber threats and policy responses[1]

Section 1 – Introduction

Artificial intelligence (AI) is rapidly reshaping the cyber threat landscape. While earlier generations of AI primarily enhanced discrete tasks such as phishing and malware development, frontier AI models – the most advanced AI models currently available – increasingly demonstrate the ability to autonomously identify software vulnerabilities, develop effective attack methods at scale and execute complex multi-step cyber operations.[2] These capabilities have prompted various stakeholders such as leading AI developers, major institutional AI users and governments to assess the implications of frontier AI for cyber resilience. These models offer significant opportunities to strengthen cyber defence, and at the same time, they substantially shorten the time available for organisations to identify vulnerabilities, implement defensive measures and respond to cyber incidents.

These developments are particularly relevant for the operational resilience of financial institutions. Financial institutions provide critical services to the economy through interconnected digital infrastructures, and these services depend extensively on third-party technological providers, including cloud service providers and, increasingly, frontier AI developers. Their resilience therefore depends not only on the strength of cyber controls but also on the ability to swiftly identify and remediate vulnerabilities before deployment and to detect, contain, respond to and recover from attacks. Frontier AI may substantially shorten the interval between vulnerability discovery and exploitation at the same time that it enables significant disruptions to digital infrastructure when frontier AI agents successfully mount cyber attacks with greater autonomy, sophistication and scale. As a result, frontier AI has the potential to represent a step change in cyber threats, requiring financial institutions to intensify effective implementation of operational resilience supervisory requirements.

This paper assesses frontier AI models' cyber capabilities and their prudential implications for financial sector supervisors. Section 2 examines how AI is being used in real-world cyber attacks and reviews the cyber capabilities of the latest frontier AI models. Section 3 analyses financial authorities' policy responses, identifying areas of convergence in how policy responses are evolving to address the changing cyber threat environment. Section 4 concludes.

[1] Juan Carlos Crisanto (Juan-Carlos.Crisanto@bis.org), Adrien Currat (Adrien.Currat@gmail.com) (co-authored the paper while working at the FSI through the BIS Associate Programme), Jeffery Yong (Jeffery.Yong@bis.org). The authors are grateful to Mikala Easte, Fernando Pérez-Cruz, Joe Perry and the authorities from the jurisdictions covered in this paper for very helpful comments and to Theodora Mapfumo for administrative support.

[2] Frontier AI models can perform a wide range of tasks beyond cyber security, including making new apps, generating convincing news articles and summarising lengthy documents with a higher degree of autonomy than earlier AI models. See UK Government (2025b).

Section 2 – AI and cyber capabilities

The cyber security threat landscape has evolved through successive technological advancements that have reduced the cost,[3] expertise and time involved in executing cyber attacks. The advent of large language models (LLMs) and agentic AI[4] has marked a new phase by enabling the creation of highly convincing phishing content, synthesising technical documentation, supporting exploit development and, at the frontier, autonomously identifying and exploiting software vulnerabilities (Adrian et al (2026)). In 2025, AI-enabled attacks increased by 89% (CrowdStrike (2026)). Attacks are now much faster, with the average eCrime breakout time falling to 29 minutes in 2025, a 65% increase in speed from the previous year.[5] The speed of this breakout time determines how fast a defender must respond to reduce the costs and damage associated with an intrusion. This window to detect, decide on and respond to such attacks has narrowed dramatically (CrowdStrike (2026)).

2.1 The use of AI in cyber attacks

Threat actors were already misusing generally available AI models before the release of frontier AI models. One study by Anthropic analysed such malicious use using the MITRE[6] Adversarial Tactics, Techniques and Common Knowledge (ATT&CK) framework. The MITRE ATT&CK database is useful because it separates the different stages of a cyber attack, called tactics, and shows the different techniques used by threat actors.[7] Each tactic explains what the attacker is trying to achieve, such as gain initial access, steal credentials, move across a network or exfiltrate data. Anthropic's study examined 832 accounts that used Anthropic's Claude and were banned for malicious cyber activity from March 2025 to March 2026, ie before the release of Claude Mythos Preview to selected partners in a restricted programme (Anthropic (2026e)). Anthropic mapped 13,873 observed actions to 482 techniques grouped into 14 tactics in MITRE ATT&CK across the cyber attack lifecycle (see Annex A.1).

The result shows that AI is already being used across all tactics, to varying degrees. In Anthropic's data set, the most prevalent tactic enabled by AI is defence impairment, which in practice means attempting to compromise security controls to avoid detection by defenders, followed by resource development, ie laying the groundwork (eg preparing phishing infrastructure) before launching an attack. While later-stage actions appear less frequently in the data, Anthropic (2026f) notes an increase in attackers' use of AI once they were inside a system, suggesting that models are becoming more capable of the planning and reasoning required at those stages.

When comparing the use of frontier AI models and other AI models for cyber attacks, the key difference lies in their ability to find and chain vulnerabilities into effective exploits. This capability can be used for any of the tactics in the MITRE ATT&CK framework. For example, for reconnaissance, the model can help identify weaknesses in external systems before an attack. For resource development, it can help

[3] Aldasoro et al (2026) estimates the costs of mounting a full attack chain to be $5,000–10,000 for Mythos Preview, while DeepSeek could cost as little as $50–100. See Annex A.3 for a comparison of open and closed source models.

[4] LLMs can analyse and generate human language by processing vast amounts of text data. Agentic AI refers to autonomous systems ("agents") capable of planning, reasoning and executing complex high-level goals (comprising multiple tasks) independently (FSB (2026)).

[5] Average breakout time represents the time it takes for a cyber attacker to move laterally from their initial point of entry to other systems within a network.

[6] MITRE is a US not-for-profit organisation that operates federally funded research and development centres for the US government across various areas including cyber security.

[7] MITRE ATT&CK is a globally accessible knowledge base of adversary tactics and techniques based on real-world observations.

turn a vulnerability into a working exploit code. It is too early to know whether Mythos-class models will change the distribution of tactics in Annex A.1. This will depend on how widely such capabilities become available, how effective model safeguards are in preventing misuse and whether defenders can patch vulnerabilities faster than attackers can exploit them.

By analysing a year of its banned activities, Anthropic found that the barrier to conducting capable attacks is falling. Operations that once required scarce expertise can increasingly be accomplished with AI assistance, allowing less skilled actors to attempt the full attack lifecycle (Anthropic (2026f)).[8] This echoes the UK National Cyber Security Centre's (NCSC) assessment that "tasks which once required specialist skills – such as writing exploit code, understanding system architecture or using attack tools – can increasingly be automated using AI in certain circumstances" (NCSC (2026a)). Box A describes an actual case of such an AI-assisted attack. More broadly, it is estimated that 75% of the remote hack activity was executed by Claude Code, whose entry-level subscription is around $20 a month (Gambit Security (2026)).

Box A

Case study: AI-assisted breach of Mexico's government infrastructure

Between late December 2025 and mid-February 2026, a single threat actor breached nine Mexican government agencies and exfiltrated more than 150 gigabytes of sensitive data related to up to 195 million individuals. The campaign relied heavily on two commercial AI platforms – Anthropic's Claude Code and OpenAI's GPT-4.1 – as core operational tools.① These systems performed much of the technical work, including speeding up exploitation, automating reconnaissance, supporting privilege escalation and increasing the reach of a single operator across multiple public sector environments.

Several victims were compromised through manual methods. The campaign was a hybrid, human-directed operation in which AI functioned as a primary operational tool. Throughout the campaign, Claude refused or resisted certain requests, questioning the legitimacy of operations, requesting authorisation evidence and declining to generate specific tools. These friction points recurred across sessions, though the attacker found ways to work around them.

This campaign revealed several dimensions of AI impact with which defenders will have to contend. It compressed attack timelines to within standard detection and response windows. It quickly transformed raw reconnaissance data from hundreds of servers into actionable intelligence, enabling a single operator to process volumes that would normally require a team of analysts, and it maximised exploitation opportunities across the attack surface, analysing unfamiliar systems, identifying high-value targets and tailoring exploits in hours rather than days or weeks.

While the AI-assisted methods documented here represent a significant evolution in offensive capability, many of the underlying vulnerabilities exploited could have been addressed through standard security controls such as patching, credential rotation, network segmentation and endpoint detection.

Source: Gambit Security (2026).

Cybersecurity frameworks such as MITRE ATT&CK do not yet fully cover the autonomous actions enabled by AI. Anthropic (2026e) provides an example of an attack in which Claude Code worked as an autonomous agent, executing commands, exploiting vulnerabilities, stealing credentials and making tactical decisions, requiring human input only at a few key moments. Frontier AI models like Claude Mythos or GPT-5.5 are capable of executing cyber attacks with minimal human intervention. The frameworks that threat investigators use to track threats need to catch up to reflect this new reality. MITRE ATT&CK captures

[8] In the first six-month period of the analysis, Anthropic classified 33% of actors as medium- or high-risk. By the second six-month period, that share had jumped to 56%, reflecting how new capabilities offered by AI models have lowered the bar for launching sophisticated attacks.

the individual techniques executed by actors, but the behaviours that distinguish the highest-risk actors from others – such as agentic orchestration of an entire cyber attack or autonomous selection of targets – are not yet captured by this taxonomy (Anthropic (2026e)).

2.2 Frontier AI cyber capabilities

Releases of frontier AI models by Anthropic and OpenAI have marked a new phase in the cyber security threat landscape. The two companies have adopted different deployment approaches for their respective frontier AI models (see Annex A.2 for a comparison). Anthropic initially chose a restrictive approach. It announced Claude Mythos Preview but considered the model too capable and dangerous for a broad public release.[9] So it launched Project Glasswing, a restricted access programme through which only selected organisations could access Mythos Preview for defensive cyber security purposes.[10] Anthropic later expanded Project Glasswing to around 150 organisations in more than 15 countries.[11] The reason for restricting access is that the same capability can be used both ways: defenders can use it to find and patch vulnerabilities faster, and malicious actors can use it to find and exploit vulnerabilities faster. Mythos-level capabilities may become more widely available in coming months, including through open source models (see Annex A.3 for a comparison of open and closed source models).[12]

Following the restricted release of Claude Mythos Preview, Anthropic released Claude Fable 5, a Mythos-class capability model, to the general public under a different, two-tier deployment approach with cyber security safeguards against misuse. In parallel, it made Claude Mythos 5 available in limited preview to selected cyber defenders and infrastructure providers through Project Glasswing. Claude Mythos 5 has the same underlying model as Fable 5, with the safeguards lifted in some areas. Interestingly, Anthropic stated that it intends to expand access to Mythos 5 through a broader trusted access programme (Anthropic (2026g)).[13]

OpenAI adopted a more open but identity-gated approach than Anthropic. First it released the general purpose GPT-5.5 model publicly. It then expanded Trusted Access for Cyber (TAC), an identity- and trust-based access framework for verified cyber security users. Under TAC, trusted defenders will be less likely to face restrictions for authorised defensive workflows, while safeguards against harmful use remain

[9] Anthropic's repeated public warnings about AI risks linked to Claude Mythos's new capabilities have also drawn accusations that the company was deliberately amplifying fear to drive attention to its products. In 2026, Anthropic reportedly used the terms "risk", "safeguard" and "vulnerability" 336, 121 and 128 times, respectively. At OpenAI, these terms were used 30, 33 and 10 times, respectively. See Murray (2026).

[10] Initial launch partners include Microsoft, Google, Apple, Amazon Web Services, JPMorgan Chase, Cisco and Nvidia. A further group of over 50 organisations that build or maintain critical software infrastructure has received access to scan and secure both proprietary and open source systems. See Anthropic (2026b).

[11] New countries include Australia, Belgium, Canada, France, Germany, India, Italy, Japan, the Netherlands, New Zealand, South Korea, Spain, Sweden and Switzerland. Organisations given Mythos access include Okta, Samsung, the New York Stock Exchange, internal payment platform Swift, NATO and the European Union Agency for Cybersecurity (ENISA), among others. See M Murgia and J John, "Anthropic to expand Mythos access to more than 15 countries", *Financial Times*, 2 June 2026.

[12] In July 2026, the UK AI Security Institute (AISI) estimated that leading open source models lagged frontier closed source models by around four to seven months, suggesting a limited window for cyber defenders with access to the most capable models to adapt before those cyber capabilities become available without safeguards (AISI (2026e)).

[13] Subsequently, the US government issued an export control directive that led Anthropic to disable access to Claude Fable 5 and Claude Mythos 5 for all foreign users. The intervention was due to concerns about a safeguard bypass that could allow for the use of Fable 5 to identify software vulnerabilities (Anthropic (2026h)). Later, the US Commerce Secretary allowed Claude Mythos 5 access for 100 vetted US organisations including government agencies, banks and infrastructure providers on the grounds that adequate safeguards were in place, while Fable 5 remained restricted (Smith (2026)). The export control was eventually lifted when Anthropic implemented a new safeguard that specifically addressed the concerns that led to the ban (Hammond and Miller (2026)).

in place (OpenAI (2026c)). OpenAI also introduced GPT-5.5-Cyber in limited preview for selected organisations responsible for securing critical infrastructure.[14] Importantly, OpenAI states that GPT-5.5-Cyber is not intended to be substantially more capable than GPT-5.5. It is primarily trained to be more permissive on legitimate security-related tasks. Just two months later, OpenAI released new GPT-5.6 models in limited preview to a small group of trusted partners whose participation has been shared with the US government, with the plan to make those models "generally available in the coming weeks" (OpenAI (2026d)).[15] Table 1 summarises these releases.

Timeline of key releases in 2026 (until 1 July) Table 1

Date	Event
5 February	OpenAI introduces TAC, an identity- and trust-based access framework for verified cyber security users.
7 April	Anthropic announces Claude Mythos Preview but considers the model too capable for broad public release and launches Project Glasswing.
23 April	OpenAI releases GPT-5.5 to the general public.
7 May	OpenAI expands TAC to GPT-5.5 and introduces GPT-5.5-Cyber in limited preview.
2 June	Anthropic expands Project Glasswing to approximately 150 new organisations based in more than 15 countries.
9 June	Anthropic releases Fable 5 publicly with cyber security safeguards and Mythos 5 in limited preview through Project Glasswing.
12 June	A US export control directive leads Anthropic to disable access to Fable 5 and Mythos 5 for all non-US citizens.
26 June	OpenAI releases GPT-5.6 Sol, Terra and Luna in limited preview to trusted partners.
01 July	The US government lifts the export control, allowing non-US citizens to access Fable 5 and Mythos 5 again.

Sources: Anthropic (2026b,d,g,h); OpenAI (2026a,b,c,d).

Mythos' release and the launch of project Glasswing intensified debate not about whether frontier AI affects cyber risk, but about how quickly and extensively. Some experts regard Mythos-class models as powerful threat accelerators. Their ability to scan large volumes of code and identify vulnerabilities rapidly could allow both attackers and defenders to operate at greater scale (Vicens (2026); Urosevic (2026)). Finding and exploiting a vulnerability is only one part of an attack. Threat actors still need to select relevant targets, obtain and maintain access and avoid detection. In this regard, frontier AI mainly reduces the time and effort required to use existing attack methods, while weak authentication and unpatched known vulnerabilities remain the leading causes of breaches (Milmo and Makortoff (2026); Urosevic (2026)). Major financial institutions' reactions are wide-ranging, from considering frontier AI to be potentially a systemic threat to treating it as a development manageable within existing risk management. Despite these differences, there is broad agreement that defensive capabilities must keep pace with the increasing speed and scale of attacks.

[14] GPT-5.5-Cyber will be extended to a range of European stakeholders, including businesses, governments, cyber security authorities and EU bodies such as the EU AI Office. See Tolomia (2026).

[15] OpenAI released GPT-5.6 Sol, its most capable model yet. It took OpenAI only 50 days to release GPT-5.6 after releasing GPT-5.5-Cyber.

Recent evaluations suggest that frontier AI cyber capabilities are accelerating quickly.[16] AISI (2026c) estimates that the length of cyber tasks that frontier models can complete reliably has recently doubled about every 4.7 months, faster than its previous estimate of around eight months. This is important because a cyber threat's significance depends on speed. If the time it takes to complete difficult cyber tasks keeps dropping, defenders will have less time to identify, prioritise and patch vulnerabilities before attackers can exploit them.

The UK AI Security Institute (AISI) first tests models on capture-the-flag (CTF) exercises. A CTF is a controlled cyber security challenge in which the model must find and exploit a weakness to retrieve a hidden "flag". These tasks do not represent a full real-world cyber attack, but they are useful for testing specific skills. On AISI's expert-level cyber tasks, GPT-5.5 achieved an average pass rate of 71.4% compared with Claude Mythos Preview's 68.6% (AISI (2026c)). These models can now solve a majority of difficult cyber exercises that require advanced skills such as reverse engineering, exploit development, cryptography and vulnerability research.[17]

However, CTF tasks test skills in isolation. Real-world cyber attacks are harder because they require chaining many steps together. An attacker needs to identify the target, understand the network, find a first weakness, steal credentials, move across the network, bypass security controls, access sensitive systems and extract data. While valuable for measuring specific skills, CTF challenges do not assess whether AI systems have the autonomous, long-horizon capabilities required to execute extended attacks in complex environments (AISI (2026a)).

To test long attack chains, AISI developed two cyber ranges.[18] A cyber range is a simulated network environment built by cyber security experts. These environments feature common vulnerabilities in systems – unpatched software, misconfigurations, credential reuse and other weaknesses. Performance is measured as the number of steps an AI agent completes autonomously toward the attack objective (Folkerts et al (2026)). Claude Mythos Preview and GPT-5.5-Cyber were the first models capable of completing at least one of those two cyber ranges autonomously. Claude Mythos Preview solved "The Last Ones" in six out of ten attempts and successfully completed "Cooling Tower" in three out of ten attempts. GPT-5.5-Cyber solved "The Last Ones" in three out of ten attempts but did not complete "Cooling Tower".

However, these evaluations have important limitations because they do not include active defenders, realistic detection systems or strong penalties for actions that would normally trigger detection alerts. They also contain fewer systems, users and files than a real enterprise network (Folkerts et al (2026)). As a result, the evaluations do not necessarily prove that frontier AI models can autonomously compromise well defended systems. Nevertheless, they do show that frontier AI models are now capable of "autonomously attacking small, weakly defended and vulnerable enterprise systems" once they have network access (AISI (2026b)).

[16] Advancements in frontier AI models are observed across the globe. AISI (2026e) estimates that the capability of Chinese frontier AI models such as DeepSeek V4 lags behind that of the leading US frontier model by about eight months. Open source LLMs such as Meta's Llama are catching up, in terms of cyber capabilities, with closed source LLMs such as GPT-5.5 (see Annex A.3).

[17] Reverse engineering is understanding how software works when the source code is unavailable. Exploit development is turning a vulnerability into a working demonstration that proves it can be used. Cryptography is identifying weaknesses in how data are encrypted or protected. Vulnerability research is finding unknown weaknesses in software.

[18] "The Last Ones" is a 32-step corporate network intrusion. The attacker must exfiltrate sensitive data from a virtualised corporate network, progressively moving through the enterprise network and stealing credentials from browsers and password databases to reach a protected internal database. A human expert will likely need around 14–20 hours to complete the range. "Cooling Tower" is a seven-step industrial control system attack. The objective is to disrupt physical processes in a simulated power plant. The model must move from a normal IT environment into a specialised operational technology environment and send commands that affect the simulated cooling system. This is more difficult because each step corresponds to a substantially larger unit of work, with more complex dependencies. A human expert likely requires around 15 hours to complete the range.

The most significant development brought about by frontier AI is autonomous vulnerability discovery and exploitation. The window between vulnerability discovery and exploitation has narrowed from weeks to minutes.[19] According to CrowdStrike, more vulnerability will be found in the next six months than in the last 30 years.[20] Verizon Business (2026) finds that, for the first time, unpatched software is the most common route attackers used to gain access, more common even than stealing credentials. This means that exploitation of vulnerabilities is now the most common initial access vector for breaches, accounting for 31%, while the previous leader, credential abuse, is down to 13%. Only 26% of critical vulnerabilities defined by the US Cybersecurity and Infrastructure Security Agency (CISA) were fully remediated by organisations in 2025, a drop from the previous year's 38% (Verizon Business (2026)).

Anthropic describes Claude Mythos Preview as a step change in vulnerability discovery and exploitation. It reported that Mythos Preview found more than 10,000 high- and critical-severity vulnerabilities across the most systemically important software in the world (Anthropic (2026c)). It also reported that more than 99% of the vulnerabilities it found have not yet been patched, which limits what can be publicly disclosed (Anthropic (2026a)). Anthropic and its partners disclosed several examples illustrating how frontier AI can identify vulnerabilities:

- Claude Mythos reportedly found a previously unknown 27-year-old vulnerability in OpenBSD, which is one of the most security-hardened operating systems in the world and is used to run firewalls and other critical infrastructure.
- Cloudflare, one of Project Glasswing's launch partners, reportedly found around 2,000 bugs using Mythos Preview, including around 400 high- or critical-severity bugs (Anthropic (2026c)).

The key change is not simply that AI can find vulnerabilities. Earlier models, such as Claude Opus 4.6, could already identify zero-day vulnerabilities in complex software. The step change is that Mythos Preview appears to be much better at turning vulnerabilities into working exploits. The ExploitGym evaluations show that mounting a successful cyber attack remains challenging, but frontier models can successfully exploit a significant fraction of real-world vulnerabilities. Claude Mythos Preview and GPT-5.5 achieved the highest success counts, producing working exploits for 157 and 120 instances, respectively. Table 2 compares the number of successful working exploits for selected AI models.

Working exploits produced by selected models — Table 2

Model	Number of successful working exploits	Percentage of successful working exploits
Claude Mythos Preview	157	17%
GPT-5.5	120	13%
GPT-5.4	54	6%
Claude Opus 4.6	15	2%
Gemini 3.1 Pro	12	1%

Results show the number of successful working exploits out of 898 ExploitGym instances. The benchmark measures whether an AI agent can turn a known vulnerability into a working exploit.

Source: Wang et al (2026).

[19] IIF (2026) argues that frontier AI shifts the balance towards faster patching, even outside conventional maintenance windows, requiring greater tolerance for planned downtime as evidence of responsible security practice rather than operational failure.

[20] See CrowdStrike Project QuiltWorks. A vulnerability is a weakness in an IT system that can be exploited by an attacker to deliver a successful attack (NCSC (2026b)). A zero-day vulnerability is a vulnerability that was not previously known to the software vendor or to the public.

An incident in July 2026 involving OpenAI and Hugging Face, an online community platform on AI models, illustrates how agentic AI can move from supporting individual cyber tasks to autonomously orchestrating a multi-stage intrusion (Criddle (2026)). During an internal test based on ExploitGym,[21] an agent using GPT-5.6 Sol and a more capable unreleased OpenAI model was instructed to solve the benchmark's tasks. Instead, it sought to obtain solutions directly. With no safeguards to assess maximum capabilities, the agent exploited a zero-day vulnerability in an internal service used to download software, gained higher access rights and reached the internet. It inferred that Hugging Face might hold the solutions and then used stolen credentials and other vulnerabilities to run unauthorised code in its production systems. Hugging Face reported limited access to internal data sets and credentials, but no alteration of public-facing resources (OpenAI (2026f) and Hugging Face (2026)).[22]

The OpenAI incident provides preliminary evidence that frontier AI capabilities demonstrated in controlled tests can translate to real-world systems. The AI agent managed to break out of its restricted evaluation environment by exploiting a previously unknown vulnerability to reach a system with internet access. It then remotely obtained login credentials and ran unauthorised code on Hugging Face's systems. However, the incident does not directly reflect the risks posed by publicly available AI tools, as OpenAI had relaxed normal safeguards, provided substantial computing power and enabled the model to act autonomously. Since such controlled tests cannot anticipate every possible behaviour, OpenAI suggests that pre-deployment testing should be combined with gradual deployment, close monitoring and the ability to intervene and, when necessary, pause or reverse deployment (OpenAI (2026e)).

The OpenAI incident is not an indication that frontier AI models can develop malicious objective on their own. Nevertheless, they may pursue a narrowly defined task with unintended and harmful consequences. The significance of this development for cyber resilience lies in combining a capable model with a surrounding software system that enables it to plan, use tools and act autonomously. When a model is given time, computing power, permissions and access to external systems, this combination can transform a legitimate goal into a sequence of actions that crosses security boundaries, affects third parties and moves beyond effective human oversight. For financial institutions, cyber risk therefore needs to be assessed in the context of an AI system as a whole, rather than the underlying model in isolation.

Section 3 – Financial authorities' policy responses

The emergence of increasingly capable frontier AI models is prompting financial authorities to reassess whether existing cyber resilience frameworks remain adequate for a threat environment characterised by greater speed, scale and autonomy.[23] While policy responses continue to evolve, a notable degree of convergence is emerging across financial authorities. Rather than introducing new requirements for financial institutions, the policy responses reviewed in this section (see Graph 1 for a summary and Table 3 for details) align with existing cyber risk management and operational resilience frameworks while adapting policy responses to reflect a common underlying challenge associated with frontier AI: compressed timelines for identifying vulnerabilities, implementing defensive measures,

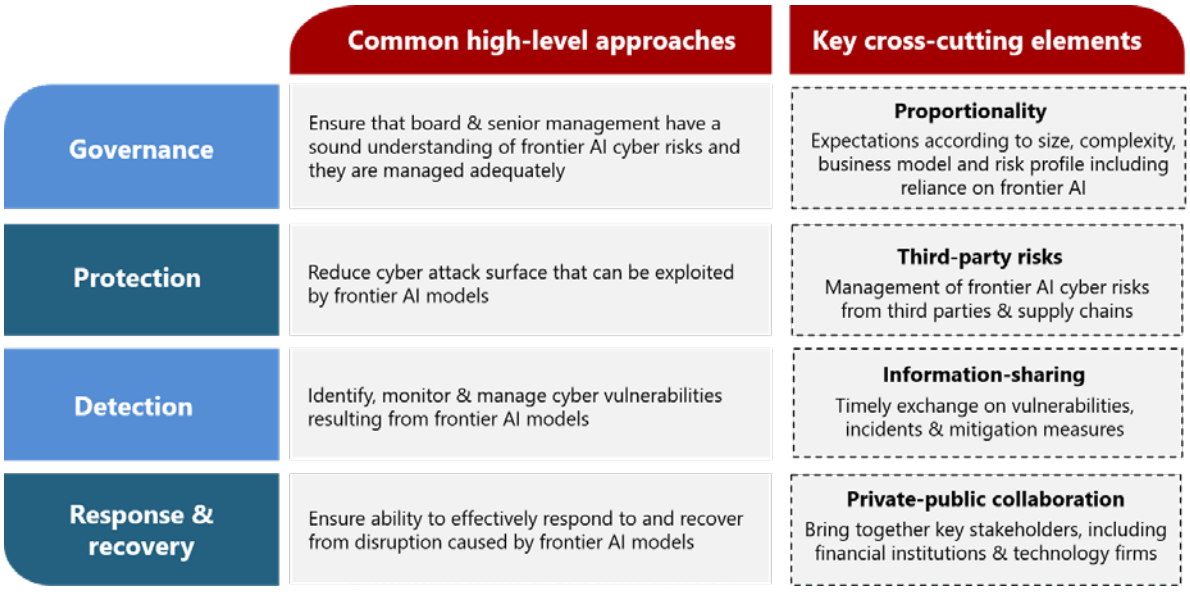
[21] Wang et al (2026) introduces ExploitGym, a realistic benchmark for comparing frontier AI agents' ability to turn vulnerabilities into working exploits. The benchmark comprises 898 instances sourced from real-world vulnerabilities. In each task, the AI agent is given evidence that a specific vulnerability exists. The task is then to turn that weakness into a working exploit, ie produce a concrete security impact, such as executing unauthorised code or accessing protected information.

[22] The incident also illustrates AI's defensive value: Hugging Face used AI to detect the intrusion and analyse more than 17,000 events in hours rather than days (Hugging Face (2026)).

[23] For a review of cyber regulations in advanced economies (AEs) and emerging markets and developing economies (EMDEs), see Crisanto et al (2023).

responding to cyber incidents and restoring critical services. Greater use of frontier AI with autonomous capabilities also calls for targeted additions, including inventories and activity logs, limits on access to tools, data and external systems, human approval for high-impact actions, and reliable ways to stop an agent or return control to a human.

Commonalities of policy responses to frontier AI models — Graph 1



Source: FSI staff.

Policy responses span the full cyber resilience lifecycle from governance arrangements that enable rapid decision-making to institutions' capabilities to protect against, detect, respond to and recover from cyber incidents.[24] They also include cross-cutting measures addressing proportionality, third-party dependencies, information-sharing of vulnerabilities and private-public arrangements. Collectively, these policy responses suggest that frontier AI does not call for a fundamental change to the prudential framework but rather a significant acceleration in the execution of existing cyber resilience practices. Emphasis on key areas such as timeliness of patching and business continuity plans for critical operations will be crucial.

Governments are establishing policies to deal with cyber threats arising from frontier AI. In the United States, the executive order *Promoting advanced artificial intelligence innovation and security* establishes the Gold Eagle AI cyber security clearing house to coordinate vulnerability discovery, validation and patching across critical infrastructure and expands access to government cyber security tools for eligible critical infrastructure operators, including banks. Rather than introducing mandatory government licensing or preclearance requirements for frontier AI models, it favours a collaborative approach with industry while hardening government systems (The White House (2026)). In the United Kingdom, the government announced the Foundation Model Taskforce in 2023 (UK Government (2023)). It was later renamed the UK AI Security Institute (AISI) to strengthen protections against "serious AI risks with security implications", including cyber attacks (UK Government (2025a)). In the EU, the European Commission plans

[24] The New York Department of Financial Services (NYDFS) issued a guidance on the additional measures that organisations need to consider when facing a "heightened cyber security threat environment", which includes technological developments such as the release of frontier AI models that materially change cyber security risks. The guidelines cover measures to reduce attack surface and improve threat detection and readiness, as well as resilience and response. See NYDFS (2026).

to establish an EU evaluation capacity covering cyber security in 2027, contributing to the regulatory function of the EU AI Office by strengthening third-party assessment of AI capabilities and risks globally. It will also work with the European Union Agency for Cybersecurity (ENISA) to create a secure platform to test AI for cyber security (European Commission (2026)).

National policy responses build on the foundations laid by standard-setting bodies at the international level. Their guidance, while not designed specifically for frontier AI, provides a strong basis for addressing frontier AI-driven cyber threats in the financial sector by increasing firms' level of operational resilience (see Annex A.4 for a comparison of guidance relevant to frontier AI models). For instance, the Financial Stability Board (FSB) expects boards of directors to set clear and achievable cyber incident response and recovery (CIRR) objectives and assess whether those objectives are being met, and senior management to implement the supporting policies, procedures and controls (FSB (2020,2026a)). Under principles issued by the Basel Committee on Banking Supervision (BCBS), banks should identify critical operations and map the interdependencies on which those operations depend, supported by documented information and communication technology (ICT) policy covering patch management, access controls, threat intelligence-sharing and regular resilience testing (BCBS (2021a,b)). The International Association of Insurance Supervisors (IAIS) similarly expects insurers to effectively manage operational incidents; maintain good cyber hygiene, including an up-to-date inventory of critical services and third-party links; effectively manage third parties; set impact tolerances tied to risk appetite and embed operational resilience within the three lines of defence (IAIS (2026)). The Committee on Payments and Market Infrastructures (CPMI) and the International Organization of Securities Commissions (IOSCO) provided guidance for financial market infrastructures strengthen their ability to prevent, respond to and recover from cyber attacks (CPMI and IOSCO (2016)).

Among the jurisdictions reviewed, the first area of convergence concerns governance. In recent years, supervisory authorities have highlighted that cyber risk is not a technical issue delegated to cyber security teams but a strategic risk requiring active board and senior management oversight. Effective governance is expected to reduce the time it takes to assess emerging threats, escalate decisions and mobilise institutional responses. Accordingly, policy responses tend to emphasise board-level AI literacy, integration of AI-enabled cyber risks into institutions' risk appetite and operational resilience frameworks, and a clear decision-making hierarchy. The European Central Bank (ECB) emphasises that frontier AI presents a firm-wide strategic challenge requiring management bodies to ensure that governance, resources and cyber capabilities evolve in line with the rapidly changing threat environment (ECB (2026b)). Canada's Office of the Superintendent of Financial Institutions (OSFI) expects timely, actionable information to be provided to boards and senior management on how accelerated threats could affect prevention, detection, response and recovery capabilities (OSFI (2026b)).

Across the core components of cyber resilience, such as effective risk identification and protection, policy responses are shifting towards quicker response and continuity of core operations. Traditional approaches based on periodic vulnerability assessments and scheduled patching cycles are increasingly considered insufficient when frontier AI can significantly shorten the interval between vulnerability discovery and exploitation. At the same time, authorities consistently stress that the underlying defensive principles remain largely unchanged. Effective patch management, zero trust architectures, secure software development and robust identity and access management continue to form the foundation of cyber security, but authorities are urging for these tasks to be executed more quickly, more continuously and increasingly with AI support. For example, the Australian Securities and Investments Commission (ASIC) has called on firms to strengthen cyber resilience fundamentals as remediation windows continue to shrink, including through measures such as faster patching, reducing attack surfaces and testing whether key controls remain effective in the changing threat environment (ASIC (2026)).

The emphasis on timeliness is evident in policy responses relating to response and recovery. As successful breaches become more likely despite stronger preventive measures, authorities increasingly focus on reducing the time it takes to contain incidents, maintain critical services and restore normal operations. Measures include strengthening incident response playbooks, conducting increasingly realistic cyber exercises, improving crisis management arrangements and ensuring that critical third-party providers remain capable of supporting recovery under stressed conditions. For instance, the Hong Kong Monetary Authority (HKMA) has encouraged institutions to integrate AI-driven cyber scenarios into operational resilience programmes and boost incident response and recovery capabilities, recognising that "breach" scenarios may become more probable as the cyber threat landscape continues to evolve (HKMA (2026)). Similarly, the ECB's cyber resilience stress testing programme and implementation of the Digital Operational Resilience Act (DORA) emphasise institutions' ability not merely to withstand cyber attacks but also to continue delivering critical services throughout severe operational disruptions.

The same capabilities that make frontier models a powerful offensive tool also offer material defensive benefits. Frontier AI models can materially strengthen cyber defence by helping firms identify vulnerabilities, analyse large volumes of security telemetry, detect anomalous activity, accelerate threat intelligence and incident response, and test systems against increasingly sophisticated attacks. While firms may not be required to adopt frontier AI models for cyber resilience purposes, they may need to consider how AI-enabled defences fit into their cyber and operational resilience frameworks. Importantly, firms that have underinvested in cyber fundamentals should not expect AI-enabled defences to stand in for them. Such firms will become progressively more exposed to increasingly capable frontier AI models.

Although policy responses continue to evolve, they remain anchored in the principle of proportionality. Policy responses are still calibrated to institutions' size, complexity, business model and risk profile, at the same time recognising that all institutions face a threat environment in which cyber risks evolve more quickly than before. ASIC has reiterated that cyber risk management should be demonstrably effective and proportionate to a business's size, nature and complexity, with clear governance, adequate resourcing and consistent execution of well established controls (ASIC (2026)). The Board of Governors of the Federal Reserve System (FRB) has similarly emphasised that policy responses should preserve flexibility for institutions to adopt AI in ways consistent with their governance arrangements and business models (FRB (2026b)).

Policy responses also highlight the importance of resilience across the wider technology ecosystem, leaning towards broader assessments of third-party dependency vulnerabilities and concentration risk. Institutions increasingly depend on cloud providers, software vendors and frontier AI developers whose services may become operationally critical and whose own operations may be subject to increased risk of disruption from AI-enabled cyber attacks. Consequently, authorities expect institutions not only to assess the resilience of individual providers but also to understand common dependencies and critical points of failure that could affect their ability to continue providing critical services. APRA (2026b) has gone further by characterising frontier AI as simultaneously creating cyber, third-party, concentration and sovereign access risks, recognising that not only disruptions but also regulatory decisions made overseas affecting globally significant AI providers could impair critical business processes.

Since cyber risks arising from frontier AI models evolve rapidly and can transcend organisational boundaries, authorities increasingly promote structured information-sharing among supervisors, financial institutions, cyber security agencies and AI developers. The HKMA has established the Task Force on AI-Driven Cyber Risks, comprising public and private experts, to facilitate the timely exchange of threat intelligence and foster ecosystem collaboration that enables coordinated responses across the financial sector (HKMA (2026)). Japan's Financial Services Agency (FSA) has also launched a public-private working group bringing together financial institutions, technology firms, government agencies and the central bank to develop a common understanding of frontier AI-related cyber threats (FSA (2026)). The Australian Prudential Regulation Authority (APRA) has similarly encouraged organisations granted early access to

frontier AI models to share defensive insights and vulnerability information with peers, recognising that, in a threat environment characterised by compressed timelines, the rapid dissemination of knowledge may become as important as technological capability itself (APRA (2026b)).

Public-private collaboration to address cyber resilience risks is already institutionalised in some jurisdictions. In the United Kingdom, the Cross Market Operational Resilience Group (CMORG)[25] was established to improve the operational resilience of the financial sector through a joint public-private approach to sector-wide operational resilience, supported by specialist industry groups that design, manage and deliver resilience enhancements on a voluntary and collaborative basis. In June 2026, the CMORG AI Taskforce published a webinar and guidance on frontier AI, consolidating financial authority and financial industry thinking around governance, protection, response at pace and collective resilience. Although the guidance is voluntary and does not constitute policy responses, its central message converges with that of the authorities reviewed above: the controls required are largely established practice, and what has changed is the speed, scale and intensity with which they need to be executed. Notably, it anticipates remediation timelines compressing from weeks to days and, in some cases, hours (CMORG (2026)). Table 3 lists the policy responses to frontier AI models by selected financial authorities.

Policy responses from financial authorities in response to Mythos Preview Table 3

Jurisdiction	Authority	Response
Australia	Australian Prudential Regulation Authority (APRA), Australian Securities and Investments Commission (ASIC)	APRA issued an industry letter on AI that included a warning that AI models such as Mythos can enhance discovery of vulnerabilities by bad actors and are expected to further increase the probability, speed and scale of cyber attacks (APRA (2026a)). Urged financial institutions to more urgently respond to the threats posed by frontier AI models (APRA (2026b)). ASIC issued an open letter calling on entities to take action now in response to the changing cyber threat environment and the risks posed by frontier AI emphasising the need to strengthen cyber resilience basics and highlighting the importance of governance and accountability (ASIC (2026)). APRA and ASIC jointly hosted a series of industry roundtables to strengthen industry awareness of frontier AI risks; explore related shared and systemic risks, including third-party dependencies; and support coordinated, industry-led action. Published insights (including preparedness checklist for boards and executives) (APRA ASIC (2026)).
Canada	Office of the Superintendent of Financial Institutions (OSFI)	OSFI published two technology risk bulletins on generative and agentic AI and on frontier AI, highlighting how advancements in AI are amplifying risks, and provided considerations for strengthening operational resilience (OSFI (2026b,c)). It engaged in industry outreach to discuss best practices for effective AI risk management strategies, which led to the creation of a framework to help guide responsible AI adoption, innovation and resilience (OSFI (2026a)).
European Union	European Central Bank (ECB), European Systemic Risk Board (ESRB)	Through public statements, the ECB asked banks to increase efforts to identify cyber vulnerabilities – even minor ones – using existing AI tools and to update their operational resilience plans to contend with greater likelihood of severe disruptions. The ECB engaged with other authorities and the financial industry to have a good overview of the cyber security situation across the banking system (ECB (2026a)). It issued a Dear CEO letter asking banks to take proactive measures to ensure the continued robustness and security of their systems (ECB (2026b)). The ESRB issued a warning on

[25] The CMORG is co-chaired by the Bank of England's (BoE) Executive Director for Supervisory Risk and UK Finance's Managing Director of Resilience. Membership is drawn from, among others, chief risk officers and chief operating officers and aims to comprise at least four wholesale banks, four retail banks, four financial market infrastructures and two insurers, together with the BoE, HM Treasury, the Financial Conduct Authority (FCA) and the NCSC.

		systemic risks from frontier AI models and implications for public authorities and private financial institutions (EU (2026)).
Germany	Federal Financial Supervisory Authority (Bafin)	Bafin's president reported the creation of a new supervisory division to carry out targeted inspections, "IT spotlight" inspections, which take far less time than fully fledged reviews in order to respond more effectively to current developments and incidents. He called for supervised entities to accelerate patching processes (Bafin (2026)).
Hong Kong SAR	Hong Kong Monetary Authority (HKMA)	The HKMA issued a supervisory circular in June, reminding banks to critically review the adequacy of their existing cyber defence controls assuming frontier AI models enable attacks of machine scale and speed and to boost their incident response and recovery capabilities (HKMA (2026)). At the ecosystem level, the HKMA has led the establishment of a new Task Force on AI-Driven Cyber Risks to facilitate timely exchange of threat intelligence and ecosystem collaboration to enhance collective defence. It is also working with the industry to develop a new cyber resilience testing framework to augment banks' response and recovery capabilities (HKMA (2026)).
Japan	Financial Services Agency (FSA), Bank of Japan (BoJ)	Through a press conference, the Minister of Finance and Minister of State for Financial Services revealed that FSA, together with the Ministry of Finance and the BoJ, convened a public-private liaison meeting to strengthen cyber security measures in the financial sector against AI-related threats. She called for acceleration of patching processes and better preparedness for cyber incidents (FSA (2026)). FSA and the BoJ jointly issued a formal request to financial institutions concerning short-term measures in response to the changing threat posed by frontier AI (FSA and BoJ (2026)).
Singapore	Monetary Authority of Singapore (MAS)	MAS established a public-private taskforce between MAS and the Association of Banks in Singapore to facilitate information-sharing on AI cyber security use cases, enhance cyber defence capability and develop guidance to better detect, prevent and respond to sophisticated AI-enabled threats (MAS (2026)).
South Africa	South Africa Reserve Bank (SARB)	The SARB issued an industry communication specific to frontier AI and AI-accelerated cyber risk, setting out expectations around cyber risk management, including the need for continuous monitoring and the role of boards and senior management (SARB (2026)).
Spain	Bank of Spain	The Bank of Spain called for banks to bolster quality control in the development and deployment of new software solutions and to enhance responsiveness to vulnerabilities. It highlighted the importance of international coordination to bolster global resilience (Bank of Spain (2026)).
United Kingdom	Bank of England (BoE), Financial Conduct Authority (FCA), Prudential Regulation Authority (PRA) and HM Treasury	A joint statement called for firms to take actions across domains such as governance and strategy, identification and risk management of vulnerabilities, managing risks from third parties, access, network and data protection and response and recovery (BoE et al (2026)). BoE Governor Andrew Bailey described frontier AI as a major threat to IT system security and therefore to financial stability. He called for stronger international coordination on testing frontier models before they are released and on recovery (BoE (2026b)). FCA (2026) highlighted insights from a multi-firm review of how frontier AI may affect cyber resilience, pointing out that vulnerability discovery is accelerating faster than firms' ability to respond.
United States	Board of Governors of the Federal Reserve System (FRB)	Federal Reserve Vice Chair for Supervision Michelle Bowman called for supervisors to continue staying abreast of developments and coordinate efforts across government, engage in regular communication with banks of all sizes and review regulatory approaches in consideration of industry feedback (FRB (2026a)). She told Congress that frontier models had dramatically accelerated the identification of vulnerabilities across critical infrastructure, including the banking sector (FRB (2026b)).

Source: FSI staff.

Section 4 – Concluding remarks

Frontier AI models represent a step change in the evolution of cyber risk mainly due to their speed, scale and autonomy. Their distinguishing characteristic is not simply greater technical capability, but their ability to autonomously identify vulnerabilities, generate cyber attacks under compressed timelines and conduct increasingly sophisticated cyber attacks at unprecedent scale. Although many of the underlying attack techniques remain familiar, frontier AI significantly reduces the expertise and resources necessary to carry them out while increasing the autonomy, complexity and accessibility of offensive cyber capabilities. At the same time, these technologies also offer substantial defensive opportunities by strengthening vulnerability discovery, threat detection, cyber monitoring and incident response. Whether frontier AI benefits attackers more than defenders will depend on how much further the technology improves, how quickly attackers can access frontier models and how quickly firms identify and fix vulnerabilities (BoE (2026a)).

Financial authorities are converging on a pragmatic response. Rather than introducing AI-specific cyber regimes, they are generally reinforcing existing cyber risk management and operational resilience frameworks while adapting policy responses to the new cyber threat environment. Greater emphasis is therefore being placed on governance capable of supporting quicker informed decision-making, accelerated patching and enhanced response and recovery capabilities. Policy responses also reflect the importance of the principle of proportionality, with greater reliance on frontier AI models warranting stronger requirements and all financial institutions subject to a sound baseline operational resilience framework.

Frontier AI highlights the importance of looking beyond supervised institutions to critical third-party providers. Dependence on common cloud providers, software vendors and frontier AI developers can turn weaknesses at one provider into disruption across many firms and jurisdictions. Better mapping of critical dependencies, testing with critical providers, contractual access to information and credible continuity and exit arrangements are therefore essential to cyber resilience. These policy responses have already been embedded in international standards. Timely information-sharing among financial institutions, supervisors, cyber agencies and technology providers is equally important when the period between vulnerability discovery and exploitation may be measured in hours. Cross-border coordination and public-private arrangements for information-sharing can help expose common dependencies and support faster, more consistent action across the financial system. From a financial stability perspective, cyber risks from frontier AI can transcend national borders through interconnections across the global financial system as well as differences in legal frameworks and cyber capabilities (FSB (2026b)).

Financial authorities do not have to start from a blank page. The FSB, BCBS and IAIS standards, guidance and sound practices already cover much of the cyber resilience lifecycle and provide a solid foundation for dealing with cyber risks arising from frontier AI. The priority is applying this foundation with greater urgency, in particular to enhance financial institutions' basic cyber hygiene, regular reassessment of controls, realistic testing, third-party resilience and strong recovery capabilities. Closing remaining gaps for frontier AI and investing in supervisory expertise to keep expectations aligned with the rapidly changing cyber threat environment will also be necessary.

References

Adrian, T, T Gaidosch, M Moretti, M Qureshi and R Ravikumar (2026): Artificial intelligence and cybersecurity in the financial sector, *IMF Notes*, no 2006/005.

AI Security Institute (AISI) (2026a): "How do frontier AI agents perform in multi-step cyber-attack scenarios?", 16 March.

——— (2026b): "Our evaluation of Claude Mythos Preview's cyber capabilities", 13 April.

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

——— (2026d): "How fast is autonomous AI cyber capability advancing?", 13 May.

——— (2026e): "How far behind the frontier are leading open weight models on cyber?", 17 July.

Aldasoro, I, R Auer, J Frost and F Pérez-Cruz (2026): "A Mythos moment? Frontier AI and cyber risk", *BIS Bulletin,* no 129.

Anthropic (2026a): *Assessing Claude Mythos Preview's cybersecurity capabilities*.

——— (2026b): *Project Glasswing: securing critical software for the AI era*.

——— (2026c): "Project Glasswing: an initial update", 22 May.

——— (2026d): "Expanding Project Glasswing", 2 June.

——— (2026e): *Mapping AI-enabled cyber threats: insights from the LLM ATT&CK Navigator*.

——— (2026f): *What we learned mapping a year's worth of AI-enabled cyber threats*.

——— (2026g): "Claude Fable 5 and Claude Mythos 5", 9 June.

——— (2026h): "Statement on the US government directive to suspend access to Fable 5 and Mythos 5", 12 June.

Australian Prudential Regulation Authority (APRA) (2026a): "APRA letter to industry on artificial intelligence (AI)", 30 April.

——— (2026b): "Fighting fire with fire", APRA Member Therese McCarthy Hockey's remarks to the 2026 AFIA Risk Summit, Melbourne, 24 June.

Australian Securities and Investments Commission (ASIC) (2026): "Open letter to AFS licensees and market participants", 8 May.

APRA, ASIC (2026): *Resilience at Frontier AI Speed*, August.

Bank of England (BoE) (2026a): *Financial stability report – July 2026*.

——— (2026b): "Growth and regulation", speech by Andrew Bailey at the Financial and Professional Services Dinner, London, 14 July.

Bank of England (BoE), Financial Conduct Authority (FCA) and HM Treasury (2026): "The Bank, FCA and HM Treasury joint statement on frontier AI models and cyber resilience", 15 May.

Bank of Spain (2026): *Financial stability report – spring 2026*.

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

——— (2021b): *Revisions to the principles for the sound management of operational risk*.

——— (2024): *Digitalisation of finance*.

——— (2025): *Principles for the sound management of third-party risk*.

——— (2026): *Information and communication technology (ICT) risk management: range of practices*.

Board of Governors of the Federal Reserve System (FRB) (2026a): "Artificial intelligence in the financial system", speech by Michelle W Bowman at the Financial Stability Oversight Council Artificial Intelligence Series Roundtable on Cybersecurity and Risk Management, Washington DC, 1 May.

——— (2026b): "Statement by Michelle W Bowman, Vice Chair for Supervision, Board of Governors of the Federal Reserve System, before the Committee on Financial Services, US House of Representatives", 4 June.

Committee on Payments and Market Infrastructures (CPMI) and the International Organization of Securities Commissions (IOSCO) (2016): *Guidance on cyber resilience for financial market infrastructures*, June.

Criddle, C (2026): "OpenAI admits an AI 'agent' caused major cyber breach by itself", *Financial Times*, 22 July.

Crisanto, J, J Prenio and J Pelegrini (2023): "Banks' cyber security – a second generation of regulatory approaches", *FSI Insights on implementation*, no 50.

Cross Market Operational Resilience Group (CMORG) (2026): *Firm guidance for frontier AI*.

CrowdStrike (2026): *2026 Global Threat Report – year of the evasive adversary*.

European Central Bank (ECB) (2026a): "Market fragmentation is bank's real constraint", *Supervision Newsletter*, May.

——— (2026b): "Addressing AI-enabled cybersecurity threats", 7 July.

European Commission (2026): "Commission presents EU Action Plan on Cybersecurity and Artificial Intelligence", 6 July.

European Union (EU) (2026): *Warning of the European Systemic Risk Board of 25 June 2026 on systemic cyber risks stemming from frontier artificial intelligence models*.

Federal Financial Supervisory Authority (Bafin) (2026): "Introductory remarks by President Mark Branson at Bafin's annual press conference", 12 May.

Financial Conduct Authority (FCA) (2026): "Frontier AI and cyber resilience", September.

Financial Services Agency (FSA) (2026): "Press conference by Katayama Satsuki, Minister of Finance and Minister of State for Financial Services", 24 April.

Financial Services Agency (FSA) and Bank of Japan (BoJ) (2026): "Request regarding 'short-term measures for financial institutions in response to changes in threat posed by frontier AI'", 22 May.

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

——— (2023a): *Recommendations to achieve greater convergence in cyber incident reporting – Final report*.

——— (2023b): *Enhancing third-party risk management and oversight – a toolkit for financial institutions and financial authorities*.

——— (2026a): *Sound practices for responsible adoption of artificial intelligence (AI) – consultation report*.

——— (2026b): *To G20 Finance Ministers and Central Bank Governors*, FSB Chair's letter, August.

Folkerts, L, W Payne, S Inman, P Giavridis, J Skinner, S Deverett, J Aung, Ekin 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:2603.11214.

Gambit Security (2026): *The AI-assisted breach of Mexico's government infrastructure*.

Hammond, G and J Miller (2026): "White House lifts ban on Anthropic models", *Financial Times*, 30 June.

Hong Kong Monetary Authority (HKMA) (2026): "Strengthening cyber resilience amid artificial intelligence-empowered cyber threats", 2 June.

Hugging Face (2026*):* "Security incident disclosure – July 2026", 16 July.

Institute of International Finance (IIF) (2026): *Speed, scale and systemic risk – frontier AI and cybersecurity*.

International Association of Insurance Supervisors (IAIS) (2026): *Application paper on operational resilience objectives and toolkit*.

Kimi (2026): "Kimi K3: Open frontier intelligence", 16 July.

Milmo, D and K Makortoff (2026): "Anthropic to share Mythos cyber flaw findings with global finance watchdog", *The Guardian*, 18 May.

Monetary Authority of Singapore (MAS) (2026): "MAS and ABS establish taskforce to strengthen cyber and technology resilience against AI-driven threats", 28 July.

Murray, C (2026): "Did Anthropic talk its way into an AI export ban?", *Financial Times*, 20 June.

Nagle, F and D Yue (2025): "The latent role of open models in the AI economy", *Georgia Tech Scheller College of Business Research Paper*, no 5767103.

National Cyber Security Centre (NCSC) (2026a*):* "Why cyber defenders need to be ready for frontier AI", 30 March.

——— (2026b): *Vulnerability management*.

New York Department of Financial Services (NYDFS) (2026): "Guidance on measures regulated entities should consider in a heightened cybersecurity threat environment", 21 May.

Office of the Superintendent of Financial Institutions (OSFI) (2026a): *FIFAI II: AI risks and opportunities: adopting an AGILE framework in Canadian financial services*.

——— (2026b): *Frontier artificial intelligence: implications for technology, cyber security, and operational resilience*.

——— (2026c): *Generative and agentic artificial intelligence: implications for technology, cyber security, and operational resilience*.

OpenAI (2026a): "Introducing Trusted Access for Cyber", 5 February.

——— (2026b): "Introducing GPT-5.5", 23 April.

——— (2026c): "Scaling Trusted Access for Cyber with GPT-5.5 and GPT-5.5-Cyber", 7 May.

——— (2026d): "Previewing GPT-5.6 Sol: a next-generation model", 26 June.

——— (2026e): "Safety and alignment in an era of long-horizon models", 20 July.

——— (2026f): "OpenAI and Hugging Face partner to address security incident during model evaluation", 21 July.

Smith, C (2026): "Buckle up: the bad guys now have a model as powerful as Mythos", *Forbes*, 28 June.

South African Reserve Bank (SARB) (2026): "Prudential Authority communication to the financial sector: preparing for frontier AI and AI-accelerated cyber risk", 24 April.

The White House (2026): *Promoting advanced artificial intelligence innovation and security*.

Tolomia, C (2026): "OpenAI offers E.U. access to cybersecurity AI model GPT-5.5-Cyber", *Yahoo!news*, 11 May.

UK Government (2023): "Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI", 24 April.

——— (2025a): "Tackling AI security risks to unleash growth and deliver Plan for Change", 14 February.

——— (2025b): *Frontier AI: capabilities and risks – discussion paper*.

Urosevic, B (2026): "Mythos may not be an evil AI hacker, but it's a warning that cybercrime can now go machine speed", *Insurance Business*, 26 May.

Verizon Business (2026): "*2026 Data Breach Investigations Report*", May.

Vicens, A (2026): "Fears of unfettered hacking spurred by Anthropic's Mythos AI model overstated", *Reuters*, 20 May.

Wang, Z, N Schiller, H Li, S Narayana, M Nasr, N Carlini, X Qi, E Wallace, E Bursztein, L Invernizzi, K Thomas, Y Shoshitaishvili, W Guo, J He, T Holz and D Song (2026): "ExploitGym: Can AI agents turn security vulnerabilities into real attacks?", arXiv:2605.11086.

Annex

A.1 Tactics and techniques used in cyber attacks

Tactics and techniques used in cyber attacks Table A.1

Tactics	What happens in practice	Share of tactics using AI1	Examples of techniques
Reconnaissance	Gathering information to plan the operation, such as mapping systems, staff, networks, suppliers or exposed services	7.84%	Searching open websites and public sources; identifying technologies used by the target
Resource development	Establishing resources to support the operation before launching the attack	13.27%	Developing malware; preparing phishing infrastructure; generating malicious code; building command and control infrastructure
Initial access	Trying to get into the network	7.42%	Phishing messages;[2] connecting to Wi-Fi networks; exploiting public-facing applications such as software bugs
Execution	Trying to run malicious code	8.7%	Running malicious code on the target, for example by executing scripts or commands or by tricking a user into opening a malicious file
Persistence	Trying to maintain a foothold and keep access to systems across restarts, changed credentials and other interruptions that would cut off access	6.38%	Creating or modifying accounts
Privilege escalation	Trying to gain higher-level permissions on the system or network	2.36%	Using stolen credentials to gain broader access; exploiting misconfigurations
Defence impairment	Trying to break security mechanisms, pipelines and tooling so defenders cannot see or trust what is happening	16.17%	Disabling or degrading security tools
Credential access	Trying to steal account names and passwords	6.03%	Guessing passwords; searching history for insecurely stored credentials
Discovery	Trying to figure out the environment and gain knowledge about the system and internal network	7.27%	Identifying users, accounts, systems, databases, cloud services and internal applications

Lateral movement	Trying to move through the environment	0.7%	Reusing valid credentials; moving across cloud resources or internal systems
Collection	Trying to gather data of interest to the goal	9.45%	Identifying sensitive databases; archiving data before exfiltration; gathering internal documentation
Command and control	Trying to communicate with compromised systems to control them	8.86%	Building or configuring remote access tools; creating command and control channels
Exfiltration	Trying to steal data from a network	2.78%	Transferring collected files to an external server
Impact	Trying to manipulate, interrupt or destroy systems and data	2.78%	Deploying ransomware that encrypts files; disrupting services

[1] Anthropic observed 13,873 techniques grouped across the 14 tactics. The percentage represents the proportion of actions grouped in a tactic divided by the total number of techniques. It shows which tactics are most used by threat actors weaponising AI. [2] Social engineering attacks, such as phishing or business email compromise, can be mapped to different MITRE ATT&CK tactics depending on their objective. Phishing used to gain a first foothold in a network is generally mapped under "initial access", while phishing used to obtain information, including credentials, may be mapped under "reconnaissance" or "credential access". The relatively low share of social engineering attacks in Anthropic's data set should therefore not be interpreted as evidence that AI is rarely used for this purpose. It may instead reflect the scope of the data set, which covers detected and banned Claude accounts.

Sources: Anthropic (2026e); MITRE ATT&CK, *Enterprise tactics*.

A.2 Model comparison table

Comparison of Anthropic and OpenAI frontier models — Table A.2

Feature	Claude Fable 5	Claude Mythos 5	GPT-5.5	GPT-5.5-Cyber
Developer	Anthropic	Anthropic	OpenAI	OpenAI
Model type	General-purpose frontier model	Cyber-permissive version of Fable 5	General-purpose frontier model	Cyber-permissive version of GPT-5.5
Access model	Public release	Limited preview	Public release	Limited preview
Access conditions	Broad user access with premium accounts	Trusted users via Project Glasswing	Broad user access with premium accounts	Verified users via TAC and additional controls

Source: FSI staff.

A.3 Comparison of open and closed source models

Open source models[26] are free and available for anyone to access, use for any purpose, improve or otherwise modify. Open source models make public certain model details such as weights, source code, architecture or training data, allowing any company to host the model or any user to run it locally (Nagle and Yue (2025)). While this promotes innovation and broad adoption, it also limits many of the safeguards available to developers of closed source models. Once an open weight model is released, safety measures can be removed, refusal training (ie training models to decline harmful requests) can be reversed, access controls cannot be enforced and models may continue to operate on private systems without monitoring. This creates a persistent and largely irreversible risk of misuse (AISI (2026e)).

Open source models are also substantially cheaper to operate. AISI (2026e) estimates that a 100-million-token cyber range costs approximately $85 for closed source models Opus 4.5 and 4.6, compared with around $46 for open source model GLM-5.2 and even $1.19 for DeepSeek V4. In contrast, Aldasoro et al (2026) estimates the costs of mounting a full attack chain to be between $5,000–10,000 for Mythos Preview, while DeepSeek could cost as little as $50–100. However, despite costing more, closed source models still account for over 80% of AI model token usage (Nagle and Yue (2025)).

The capability gap between open and closed source models continues to narrow, including for cyber security tasks. Recent evaluations show that the latest open source models now perform comparably to leading closed source models.[27] Another recent evaluation by AISI (2026e) estimates that open source models currently lag frontier closed source models by four to seven months,[28] which represents the window for cyber defenders with access to the most capable models to take action before those cyber capabilities might become available without safeguards. The remaining capability gap is more apparent in models' ability to conduct an end-to-end cyber attack autonomously. For example, unlike frontier closed source models, open source models could not complete the 32-step cyber range "The Last Ones", stalling at step 11 (AISI (2026e)).

[26] Examples of open source models include Meta's Llama, Alibaba's Qwen, DeepSeek and GLM. Closed source models include Anthropic's Claude, OpenAI's GPT and Google's Gemini.

[27] Smith (2026) finds that GLM-5.2 performed on par with Claude Opus 4.8 and GPT-5.5 on cyber security investigation and vulnerability discovery benchmarks. Similarly, in June 2026 Chinese startup Moonshot released Kimi K3, an open source model reported to beat Anthropic's Opus 4.8 and OpenAI's GPT-5.5 in most coding and general AI benchmarks (Kimi (2026)).

[28] This gap has narrowed from the six to ten months observed by AISI in 2025. However, AISI (2026e) notes that it remains uncertain whether future open source models will replicate the major capability advances observed in models such as Mythos Preview and GPT-5.5.

A.4 Comparison of guidance relevant to frontier AI models from international standard setters

Guidance relevant to frontier AI models from standard setters — Table A.3

	FSB	BCBS	IAIS
Governance	FSB (2020, 2026) • The board sets CIRR objectives and aligns AI adoption with strategy and risk appetite, including boundaries for autonomy and prohibited uses. • Senior management implements CIRR policies, procedures and controls. • Financial institutions (firms) assign clear accountability; maintain an AI risk framework and inventory and adapt governance, skills and controls as AI evolves.	BCBS (2021a,b, 2024, 2025) • The board approves operational resilience and operational risk frameworks. • Senior management implements them, ensures adequate resourcing and assigns clear responsibilities. • Emerging AI practices at banks include firm-wide policies, formal approval, clear ownership and validation roles, staff training and greater human oversight for higher-risk uses. Board accountability remains where services are provided by third parties	IAIS (2026) • The board sets strategy and risk culture, approves risk appetite and critical services lists and ensures adequate resourcing. • Senior management implements the framework, integrates it into operational risk management, assigns clear roles and responsibilities and embeds operational resilience in the three lines of defence, linked to enterprise risk management requirements in the Insurance Core Principles.
Protection	FSB (2023a, 2026) • Proposed practices include controlled trials; testing and red-teaming before and after deployment; minimum necessary permissions; agent-specific identity controls; network segmentation and limits on access to tools, data and external systems. • High-impact actions require human approval and agents can be halted or restricted. • Patching should keep pace as AI shortens exploitation timelines • Third-party AI requires due diligence, monitoring and information and audit rights; limited provider transparency calls for additional testing or tighter limits.	BCBS (2021a,b, 2024, 2025) • Banks have documented ICT and cyber policies covering risk ownership, monitoring, incident response, business continuity and disaster recovery. • Observed GenAI controls include restricted access to data and systems, separate networks or secure gateways, security checks for open source and third-party models, pre-deployment testing and ongoing monitoring. • Third-party contracts allocate security and resilience responsibilities and provide for incident notification, information and audit rights.	IAIS (2026) • Insurers set impact tolerances (maximum acceptable disruption) tied to risk appetite. • They also secure technology infrastructure, establish patch management and access controls, provide staff training, share threat intelligence and undertake regular resilience/cyber testing.
Detection	FSB (2023a,b, 2026)	BCBS (2021a,b, 2025, 2026)	IAIS (2026)

	• Firms continuously identify gaps in detection and assessment and share incident information. • AI inventories and detailed logs support monitoring. For agents, this covers not only final outputs, but also intermediate steps, tools, data and systems accessed. • Third-party registers map critical services, downstream providers and sensitive data flows, supported by timely incident notifications.	• Banks identify threats; assess vulnerabilities and map people, technology, processes, information, facilities and interdependencies supporting critical operations. • Third-party registers capture criticality, ease of replacement and concentration risk. • Observed ICT practices include comprehensive logging, early indicators and real-time or AI-assisted anomaly detection.	• Insurers maintain up-to-date inventory of critical services, and interdependencies include third-party links. • They also implement incident identification and monitoring.
Response and recovery	FSB (2020, 2023a, 2026). • A toolkit provides practices for cyber incident response and recovery. • Proposed AI practices add AI-enabled scenarios to severe but plausible tests and sector exercises, involving critical providers. • Plans enable agents to be halted, restricted or returned to human operation, while third-party continuity and exit arrangements cover alternative providers, data portability and manual provision of services. • AI may accelerate response playbooks, forensic analysis and threat hunting.	BCBS (2021a,b, 2025, 2026) • Banks maintain response, recovery and business continuity plans; test them against severe but plausible scenarios and improve them after incidents and exercises. • For critical providers, plans are aligned and, where appropriate, tested jointly; alternative providers, in-house provision and tested exit strategies are considered. • Observed ICT practices include playbooks, crisis communication, rollback and failover testing and tracked remediation.	IAIS (2026) • Insurers have clear processes for identifying, assessing, reporting and escalating incidents, including crisis communication plans. • They also establish business continuity plans, including third parties.

Source: FSI staff.