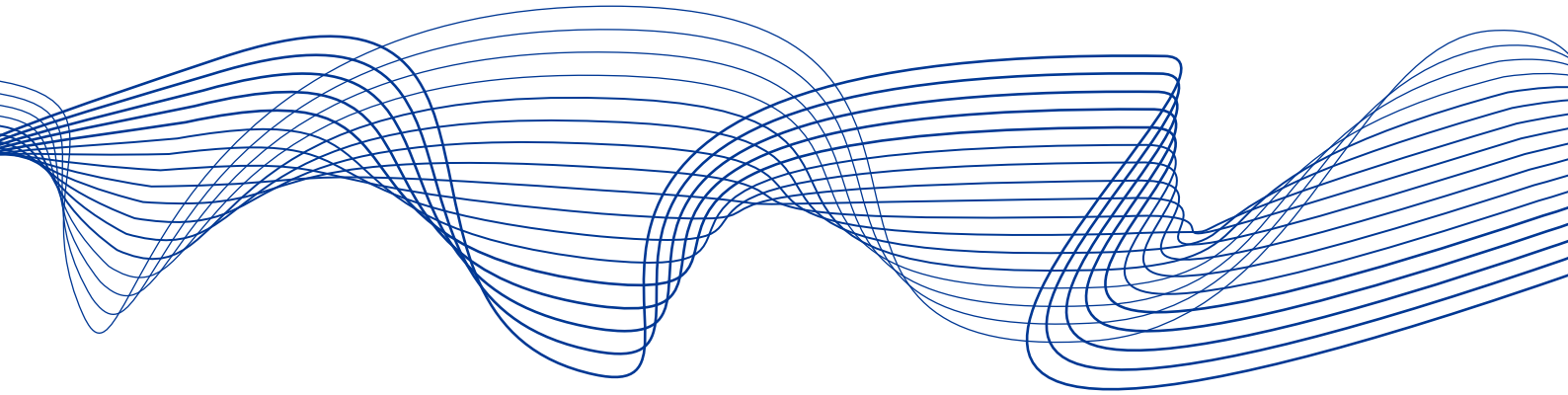
ESRB
European Systemic Risk Board
European System of Financial Supervision

Addressing frontier AI models with cyber capabilities from a financial stability perspective

July 2026



Contents

Executive summary

This note examines the emerging systemic cyber risks for the EU financial system arising from recent developments in frontier artificial intelligence models (FAIMs).[1] These developments materially change the scale and structure of the cyber risk landscape by enhancing the capabilities of both offensive and defensive actors.[2] As a result, this constitutes a structural increase in systemic cyber risk to the EU financial system for which no fully effective mitigation framework currently exists. Any effective mitigation would require coordinated efforts involving all affected parties, including AI providers, software providers, security firms, open source maintainers, financial institutions and authorities at both national and EU level.

FAIMs are reported to be highly capable in the cybersecurity domain, able to carry out fully automated[3] attacks on complex systems, including vulnerability discovery, exploit development and weaponisation. Their capabilities include the ability to autonomously identify previously unknown vulnerabilities across major operating systems and widely used software that underpin today's information and communications technology (ICT) environments and financial infrastructure.[4] FAIMs are also significantly better than previous models at finding ways to exploit vulnerabilities, outperforming earlier models in terms of efficiency, speed and accuracy. In response to the associated risks, providers have restricted the dissemination of FAIMs in the belief that unrestricted access would pose an excessive risk at this stage. Consequently, several controlled access initiatives have been introduced through which selected institutions are able to use these systems to protect critical financial infrastructure and identify and mitigate unknown vulnerabilities. Approaches vary in terms of access: in some cases, access has been highly restricted, while in others it has been structured in tiers, enabling broader availability.

The emergence of FAIMs with capabilities rivalling those of humans, together with the proliferation of FAIM-powered weaponised critical vulnerabilities, has drastically increased risks to individual critical and important functions, financial institutions and the entire financial system. There are several factors behind this, including:

---

1 For the purposes of this note, frontier AI models (FAIMs) means advanced general purpose AI models capable of materially affecting offensive or defensive cyber operations.

2 As the only two models that have been publicly announced, the report uses Anthropic's Claude Mythos and OpenAI's GPT-5.5-Cyber as examples, although it should not be interpreted as applying solely to these models, but rather to the broader inflection point in AI capabilities. The considerations presented in this report, as well as the policy options discussed, should therefore not be regarded as limited to specific models, but as applying to any type of contemporary FAIM development.

3 This does not mean that the entire attack chain is fully autonomous, but rather that FAIM agents can be used in a pipeline to carry out the attack without human intervention.

4 While there are limited data on the real-world application of FAIMs, the available evidence is broadly consistent and corroborative, indicating that FAIMs are capable of discovering large numbers of vulnerabilities, a significant proportion of which are of high or critical severity.

- a reduction in the time buffers currently relied on to ensure the continuity of financial services while patches are applied to affected systems;
- an increase in the number of discovered critical vulnerabilities in both new and legacy systems, which risks overloading current vulnerability management frameworks;
- a similar impact on common exposures through critical third-party providers, shared technological ecosystems and widely used open source components;
- a heightened need for incident response due to a potential rise in successful attacks;
- greater asymmetry in the cyber domain, with much fewer resources now needed to carry out complex cyberattacks; and
- a presumably large number of undiscovered vulnerabilities in the digital infrastructure underpinning the financial system, where FAIMs may give threat actors an advantage in the short to medium term.

These drivers affect operational resilience across four main areas: (a) time, relating to the collapse of defensive time buffers and the need to apply urgent patching without causing operational failures; (b) defenders' capabilities, in the sense of their ability to test systems and to protect, detect and respond to threats; (c) concentration, relating to dependencies on a limited number of AI providers, cloud providers, widely used software, and open source components; and (d) the capabilities of authorities, relating to calibration of expectations, stress testing and preparedness measures as FAIMs continue to develop.

Moreover, operational resilience will depend on the extent to which financial institutions have visibility over AI deployments, including a clear understanding of where and how AI is used across customer interfaces, internal agents, business processes and third-party integrations.

These capabilities are likely to increase the immediate threats facing individual ICT systems and the financial sector as a whole. While defensive capabilities are likely to improve in the long term, defenders remain structurally constrained by operational, regulatory[5] and technological factors, resulting in a decisive short to medium-term advantage for attackers.

Financial institutions are differently resourced and equipped to cope with the challenges posed by FAIMs. As a result, some institutions are more exposed to these threats, and given the interconnectedness of the financial system, this may erode trust and exposes the financial system risks.

The concentration of AI development in third countries amplifies the risks to the EU financial system. Leading AI providers used by the financial sector are, with

5. Regulatory frameworks require financial institutions to follow structured processes for change management, testing, validation and risk assessment. While these processes are needed to ensure system stability and resilience, they may also slow the deployment of defensive measures in response to rapidly evolving cyber threats.

one exception, headquartered in third-country jurisdictions. This creates strategic dependencies and geopolitical risk, potentially making access to FAIMs a source of geopolitical leverage and increasing jurisdictional asymmetries. Moreover, access to FAIMs cannot be assumed, as it may be subject to extensive export controls or other restrictions.

Overall, developments in FAIMs should be treated as a source of systemic risk. Should incidents propagate through payment systems, clearing and settlement or other operational bottlenecks, they could severely disrupt or shock the financial system, undermine public confidence and lead to heightened financial volatility. Without timely, coordinated and efficient action at EU level, the current asymmetries may evolve into structural vulnerabilities, increasing the likelihood that FAIM-induced cyber stress translates into a systemic event.

1 Introduction

The EU financial system is highly digitalised, automated and interconnected and depends on a wide array of information and communication technologies (ICTs) and related information assets. The system is reliant on this to remain both functional and efficient. However, this dependency exposes financial institutions and market infrastructures to threats that may impair specific critical or important functions, or materialise in a correlated or widespread manner, causing systemic disruption across the entire sector. Cyber risk has therefore been recognised as a source of systemic financial risk by macroprudential authorities, including the ESRB in its 2022 Report on Mitigating Cyber Risk[6] and in its 2020 Report on Systemic Cyber Risk[7].

Recent developments in frontier AI models (FAIMs) materially change the nature and magnitude of cyber risks. In spring 2026, a small number of FAIMs were reported to incorporate, among other capabilities, cyber capabilities facilitating the autonomous[8] discovery and effective exploitation of previously unknown software vulnerabilities ("zero-days"), including vulnerabilities in major operating systems and widely used software that form the backbone of today's ICT environments and financial infrastructure. At the time of writing, both AI providers that have publicly announced FAIMs with such capabilities, namely Anthropic and OpenAI, have opted not to provide their latest cyber models through unrestricted public access. Instead, they have launched controlled-access initiatives[9] ("Project Glasswing" and "Daybreak", respectively), through which selected technology firms, cybersecurity companies, financial institutions and governments can use these cutting edge models to identify and remediate vulnerabilities in critical software and infrastructure.

Official benchmarks[10] produced by the UK AI Security Institute (AISI), as shown in Appendix A, illustrate the steep increase in the cyber capabilities of current FAIMs, as well as their associated costs. Given a constant budget, the latest FAIMs are able to complete all steps of the UK AISI cyber benchmark, something that previous models were unable to accomplish. AI models released at the end of 2025, just months before the publication of this note, were, on average, able to complete only 34% of the same UK AISI cyber benchmark.

---

6 ESRB (2022), *Mitigating systemic cyber risk*, January.

7 ESRB (2020), *Systemic cyber risk*, February.

8 This does not mean that the entire attack chain is fully autonomous, but rather that AI agents can be used within a pipeline to carry out the attack without human intervention.

9 The early access initiative Project Glasswing has been effectively halted, as Anthropic has cited US government export controls, requiring access to the affected models to be suspended for all customers in order to ensure compliance.

10 The benchmark refers to the UK AISI cyber range known as "The Last Ones" (TLO): a 32-step corporate network attack simulation spanning initial reconnaissance through to complete network takeover, which we estimate would take a human operator 20 hours to complete. A more detailed description of the range can be found in a paper published by the UK AISI.

While the emergence of FAIMs has brought unprecedented cybersecurity capabilities, it is important to recognise that this was not the primary objective when training the models. Indeed, the models were initially trained to produce high-quality code rather than to identify weaknesses in it. Consequently, FAIMs that are deliberately trained to audit software and find weaknesses could presumably have even greater capabilities. This also strongly suggests that models with similar or superior capabilities will become available in the near future, marking the emergence of a broader class of AI models with advanced cybersecurity capabilities. In practice, considerations and policy options related to the latest FAIMs should therefore be understood as being illustrative of a broader class of emerging FAIMs, rather than being limited to any single provider or version.

The emergence of FAIMs with cyber capabilities has several implications of direct relevance, including:

- significantly increasing the rate at which critical vulnerabilities can be discovered in both new and legacy systems, likely overloading existing vulnerability management frameworks;
- reducing the time buffers (from weeks to hours) that are currently relied on to patch and mitigate vulnerabilities before weaponised versions become widely available;
- enabling the reverse engineering of released patches in order to exploit vulnerabilities before systems can be patched;
- reducing the human resources required to conduct complex cyber operations, thereby creating an asymmetric advantage for attackers over defenders, where defenders have traditionally benefited more from advances in AI.

In addition to the above, these developments are happening amid heightened geopolitical tension, when FAIMs are increasingly being viewed as a matter of national security by the jurisdictions in which they are developed. The current geographical concentration of leading AI providers in the United States and China[11] leaves the European Union exposed to strategic dependency and geopolitical risks. Given the capabilities of these models in the cyber domain, which is widely recognised as a domain of warfare[12] in contemporary doctrines, it is entirely conceivable that AI models with cyber capabilities will become subject to dual use technology[13] export controls, much as encryption algorithms were in the past.

Although not initially subject to export controls, access to FAIMs was, in an initial phase, granted primarily to US-based institutions. Subsequently, a limited number of jurisdictions were granted access, although neither of the two FAIM

---

11 While most leading AI providers capable of developing FAIMs are headquartered in the United States, some, such as DeepSeek, ByteDance and Alibaba, are headquartered in China. The only prominent provider headquartered in the EU is Mistral, which is based in France.

12 Recognised formally by NATO, among numerous other jurisdictions, including the United States, China, the United Kingdom and Russia.

13 In this context, dual use technology refers exclusively to technology with both civilian and military applications.

providers made the technology available to EU counterparts. While the controlled access initiatives were later expanded to include EU institutions, access did not extend to all EU Member States. However, Anthropic has since revoked access to its latest models, citing export control directives that restrict access to US nationals only.

As most FAIM providers, and any control regimes governing access to FAIMs, fall under third-country jurisdictions, early access to future FAIMs for EU institutions cannot be assumed. This further illustrates how asymmetries in access to FAIMs among EU financial institutions could become a permanent feature.

These models enhance both offensive and defensive capabilities, albeit asymmetrically. The former is likely to take the form of significantly lowered barriers to entry for various threat actors, while the latter is likely to support and develop defensive cybersecurity functions within financial institutions, specifically with regard to the early detection of vulnerabilities and attacks, and continuous remediation of vulnerabilities in complex ICT systems.

Possible ways in which the risks described in this note could materialise and escalate can be explained through hypothetical scenarios. Appendix B describes three such conceptual scenarios, which are provided for illustrative purposes only, without any implied likelihood of risk materialisation or occurrence. Moreover, the scenarios are not designed to be mutually exclusive nor exhaustive. They represent simplified, intentionally exploratory examples and may be used to highlight existing risks and mechanisms, including amplification and transmission channels. They are meant to facilitate discussion and ensure a common understanding of how, and to what extent, FAIMs could affect the financial system, rather than to provide an authoritative ladder of escalation or predictive pathways. Taken together, the three scenarios show a progression from idiosyncratic and localised events to full-scale systemic disruption, highlighting how FAIM capabilities can affect the financial system. The scenarios also show that risks can materialise in different forms, amplifying existing vulnerabilities within the financial system.

2 Impact of FAIMs on cyber risk

This section outlines the key aspects of how the proliferation of FAIM capabilities may materially alter both the scale and structure of cyber risk. It describes both changes in level of existing risks and changes in the structure of the risks that financial institutions are likely to face. It looks at both individual entities and the entire financial system.

2.1 Increased asymmetries in the financial sector

This subsection covers three sources of asymmetry in the financial sector: the first between jurisdictions, the second between defenders and attackers, and the third between better resourced and less well-equipped financial institutions. The first is largely driven by limits in terms of access to FAIMs. Jurisdictions with advanced cyber capabilities may seek to cement their technological supremacy, potentially using that position as a source of geopolitical leverage as well. The second relates primarily to the lower cost of carrying out attacks. The third is a product of the previous two and implies that the cost of responding to cyberattacks is disproportionately high for less well-equipped institutions.

From a jurisdictional perspective, FAIM development risks increasing asymmetries in capabilities across jurisdictions. The main reason for this is the aforementioned geographical concentration of prominent AI providers, which leaves the EU exposed to strategic dependency and geopolitical risk, as well as the possibility that access to FAIMs could be used as a coercion tool and as an offensive tool for applying direct pressure on the EU financial system.

Meanwhile, the asymmetry between attackers and defenders comes from the growing availability of AI models, benefitting not only defenders but also attackers. One aspect of this is that FAIMs lower the price of admission for attackers. Threat actors will increasingly be able to rely on FAIMs to conduct the more technically advanced and labour-intensive elements of their offensive operations. This includes using FAIMs automatically and rapidly to:

- scan for vulnerabilities in widely used ICT systems, including released patches for such systems;
- chain together existing, less severe vulnerabilities into more potent vectors;
- generate functional exploits;
- discover complex attack paths previously overlooked;
- weaponise exploits to match a target institution;

- move laterally once a foothold has been gained in a target entity in order to capitalise on the successful attack.

Defenders also stand to benefit, albeit less so than attackers, since many are often constrained by either:

- operational and regulatory requirements relating to, for example, uptime, change management processes, testing and validation;
- dependencies on vendors and other third-party providers, which may be slower to adopt FAIMs for patch development and deployment;
- dependencies on open source software that lacks the capacity to be updated at a pace consistent with the rate of vulnerability discovery;
- dependencies on deprecated and legacy systems, for which improving the level of security is hard to achieve because they were often not designed with security in mind; or
- a combination of these.

The availability of FAIMs will further increase the asymmetry between attackers and defenders by lowering the price of admission for attackers. Attackers previously unable to conduct advanced cyber operations will be capable of targeting specific systems or entities on a relatively low budget. This substantially lowers the threshold for what type of threat actor should be considered a qualified threat to a financial institution, thereby significantly increasing the pool of potential attackers. Overall, this points to a potential capability gap between attackers and defenders, with threat actors able to leverage FAIMs more rapidly and efficiently than the defensive functions of individual financial institutions.

This increase in asymmetry is likely to result in defenders needing increased efforts and resources to protect their services, while lowering these needs for attackers. This, in turn, will be reflected in the overall cyber risk faced by individual financial institutions, requiring them to commit additional resources to defensive capabilities.

Given the existing differences in how financial institutions are resourced and equipped to cope with the challenges posed by FAIMs, those with less favourable circumstances will likely face a disproportionate burden, generated by some costs being fixed[14] **rather than scaled, resource constraints, and limited access to specialists.** This places such institutions at a disadvantage, potentially affecting their ability to adequately fulfil their responsibilities pertaining to operational resilience and leading to an erosion of trust in the financial sector. However, there are several industry-driven initiatives in place addressing similar problems, and extending such initiatives to include the use of FAIMs is worth considering. This would potentially allow private financial institutions to respond to the threat as a community, thus relieving part of the systematic pressure.

[14] Such as licensing of defensive tooling, staffing and regulator interaction.

2.2 Accelerated vulnerability discovery

As previously noted, the discovery of critical vulnerabilities in commercially used software is relatively infrequent, labour-intensive process reliant on highly skilled specialists. This has allowed financial entities to adopt vulnerability management frameworks based on the risk-based approach, whereby vulnerabilities are prioritised according to their severity, likelihood of exploitation and impact on the organisation. However, FAIMs are reportedly able to:

- autonomously identify previously undiscovered vulnerabilities, including by chaining together less severe vulnerabilities, across all major operating systems, commercial software and other ICT systems, including systems previously considered fully secure and exhaustively tested to the gold standard; and
- discover vulnerabilities with unmatched speed and accuracy, finding flaws that have gone undetected in major systems for years.

Capabilities such as these could be detrimental to the financial system if such models were to become available to threat actors, thus enabling:

- the discovery of novel exploits in financial systems for which there are no known indicators of compromise and no readily available security patches should an attack be discovered;
- positioning within critical systems for long-term data exfiltration or large-scale coordinated disruption;
- a vast number of exploits to be generated targeting high and critical severity vulnerabilities, while also enabling attacks to be perpetrated against multiple entities at once, crippling the system;
- the systematic deployment of autonomous attacks against multiple exposed financial ICT systems, significantly raising the risk of correlated compromise;
- the overloading of existing vulnerability management frameworks, which rely heavily on risk-based prioritisation of vulnerabilities, forcing financial institutions to selectively defend against some attacks while knowingly ignoring others.

These considerations make it harder for any kind of institution operating within the financial system, whether private or public, to prioritise mitigation and remediation effectively. Doing so would require additional resources and effectively prolong exposure to vulnerabilities, leading to a short to medium-term increase in cyber risk at the individual institution level. Over a longer-term horizon, the defensive capabilities of such models will likely outweigh their offensive capabilities, as has been the case with previous technological shifts.

2.3 Collapse of defensive time buffers

Today's defensive vulnerability management practices rely on built-in time buffers. This often includes a non-mandatory but standardised ethical disclosure window, in which any vulnerability, unless otherwise agreed, is reported to the vendor before public disclosure, thus giving the vendor time to develop a remedy and distribute a patch. This period is often as long as 90 days for more severe vulnerabilities. However, even after a vulnerability becomes known to the public, it is often only presented as a proof of concept or a theoretical attack, meaning that no exploit is yet available. Traditionally, the crafting of such exploits, and their subsequent weaponisation for use against a specific organisation, has been a time-consuming process. This has historically created a time lag before a discovered vulnerability can be exploited under real life conditions, thus giving defenders a window of time in which to deploy defensive measures and prepare for a potential attack against their systems.

Conversely, the release of patches also facilitate attacks, as patches can be used as guidance on how best to attack. Given that FAIMs are already capable of rapidly creating weaponised exploits for known vulnerabilities, patches may themselves be utilised to find those vulnerabilities. Indeed, every patch can be reverse engineered, a task that even current models are consistently able to do in very short timeframes. This behaviour effectively uses the patch as a guide to which part of the code is vulnerable, thus narrowing the search significantly and creating ample opportunity to craft a working exploit. This exploit can then be used to target organisations that have yet to apply the original patch to their system, increasing the likelihood of a successful attack.

- Current FAIMs are reportedly able to compress this entire timeline into significantly less than a day, thereby shortening remediation windows and reducing the period between the first exploitation and widespread, fully automated exploitation from weeks to just hours.
- The emergence of unpatchable time windows, where an entity must develop, test and deploy a patch in a critical system faster than the patch can be reverse engineered and the underlying vulnerability exploited.
- Greater risk of incidental outages affecting critical or important functions underpinning financial services, as patching must be carried out frequently and rapidly in critical systems. This increases the likelihood of the integrity of such systems becoming compromised, thus risking outages and rollbacks.[15]

This is tantamount to a collapse of defensive time buffers, reducing the time available for response and effective defence and thereby significantly reducing the effectiveness of existing defensive protocols.

[15] The CrowdStrike incident of July 2024 goes to show the impact that disruptions in underlying solutions can have on the financial system. To date, it is regarded as the largest operational incident in history, with estimates suggesting that the total cost for Fortune 500 companies alone exceeded $5 billion.

The collapse of time buffers significantly increases the likelihood of disruptions to individual critical or important functions. This increase in cyber risk stems from the pressure placed on organisations to either patch quickly or to become prime targets for exploitation, leading to a further systematic increase in cyber risk for individual entities, with possible systemic implications depending on how the risk materialises. Ultimately, repeated stress is likely to undermine public and investor confidence, eroding trust in individual institutions and the financial sector as a whole. This is because more exposed financial institutions would no longer be perceived as secure, given their inability to defend against rapid cyberattacks and the resulting disruption to their financial services.

2.4 Compromised operational continuity

It is standard practice in the financial sector to use a risk-based approach to vulnerability management. Vulnerabilities are typically assessed according to their severity, likelihood of exploitation, expected impact on the target system, and any mitigating actions taken to reduce those properties. This leads to a manageable subset of vulnerabilities requiring immediate attention, allowing an institution to focus its defensive capabilities on those vulnerabilities while addressing lower-risk vulnerabilities over a longer timeframe.

Current vulnerability management frameworks are ill-suited to handling large numbers of high and critical severity vulnerabilities. The main reason is that, under current frameworks, each vulnerability requires an individual risk assessment that takes the target systems into account, both for prioritising mitigating actions and for assessing the risk of such efforts. This is a time-consuming process that can create bottlenecks within the framework. Moreover, if the framework is flooded with critical vulnerabilities, lower-priority vulnerabilities may remain unaddressed for prolonged periods. To address this, financial institutions will likely have to consider trade-offs between:

- patching vulnerabilities rapidly, while placing overall system stability at risk;
- assessing and testing, leaving systems exposed to known vulnerabilities;
- prioritising according to risk, in the knowledge that less severe vulnerabilities will remain unaddressed for a prolonged time and that such vulnerabilities might be chained together to form more potent vectors; and
- shifting to a different approach, carrying the risk of not remediating severe vulnerabilities at a sufficient speed.

The likely overload of incident management frameworks is analogous to the challenges associated with vulnerability management.[16] Moreover, establishing and maintaining visibility of AI deployment within financial institutions is challenging, particularly where AI is used in customer interfaces, internal agents, business

[16] A key difference is that the trade-off lies between accepting that attackers have more time inside ICT systems and allowing for an immediate and automated response that risks operational continuity.

processes and third-party integrations. Limited visibility could, in turn, amplify risk, extending not only to the internal management of an institution, but also to the response capabilities of third-party providers and authorities such as law enforcement, supervisory authorities and coordinating authorities.

To summarise, the introduction of FAIMs is likely to overload current frameworks. This strain may weaken operational resilience, increasing the likelihood of operational incidents and outages affecting critical or important functions, and forcing financial institutions to accept higher levels of operational risk.

3 Systematic and systemic risk

The emergence of FAIMs in the cyber domain will have implications that stretch beyond the risks faced by individual financial institutions. The developments described above may materially alter both the scale and structure of cyber risk, increasing both direct and indirect risks to financial stability. On top of this, systemic risk may emerge through a loss of confidence in the capacity of the financial system to withstand and absorb systematic pressure.[17] At present, no single mitigation framework appears sufficient to fully address these risks should they materialise simultaneously or across shared dependencies.

3.1 Increase in attacks on the financial community

Historically, most cyber incidents affecting the financial sector have been treated as idiosyncratic and localised events, affecting the operational status of individual institutions, but not impairing the wider system. While the impact has usually been limited to the affected institution, the stability of the financial system relies on attacks rarely succeeding and seldom overlapping. It also relies on advanced and coordinated attacks being costly to perform, both in terms of time and resources, and rarely being directed at the financial system as a whole. With FAIM-powered automated attacks that are less resource-intensive and easier to coordinate, there is a risk of both systematic and systemic impacts across the financial sector. This may result from a combination of a persistent increase in the number of severe vulnerabilities discovered and shorter remediation windows, resulting in a much higher baseline. This, in turn, forces the system to absorb a greater number of disruptions. If those disruptions become too numerous or severe to absorb, they may create cumulative effects across the system, amounting to systematic risk when multiple financial institutions are attacked independently, and to a systemic risk when such attacks are sufficient to harm the entities collectively and impair the capabilities of the EU financial system. Financial market infrastructures, together with their associated ecosystems[18], are particularly vulnerable to such systemic stress, as they are less substitutable.

3.2 Common exposures

The EU financial system depends heavily on third-party providers and open source software, giving rise to both third-party risk[19] and concentration risk. This includes dependencies, as already mentioned, on a limited number of providers

---

17 ESRB (2022), *Mitigating systemic cyber risk*, January.

18 Including market participants, financial market infrastructure (FMI) vendors and their products, and FMI service providers.

19 Including what is commonly referred to as nth-party risk, where the risk stems not from the third party, but from suppliers further down the chain.

of AI models, cloud infrastructure and cybersecurity services. Similarly, as the technology stacks used by individual financial institutions are often either the same or substantially similar, financial institutions are frequently exposed to common vulnerabilities. Any vulnerability found to exist in widely used technology could therefore lead to a clustering of successful cyberattacks against the financial sector. In the same way, vulnerabilities discovered in the technology of a third-party provider or an open source library could lead to the widespread supply-chain compromise affecting large sections of the financial system, with the coverage dependent on the market share of the provider or library. Both of these scenarios would lead to correlated disruptions across the wider financial system, constituting an increase in systemic risk.

The DORA register of information on ICT third-party arrangements should provide greater means to manage third-party risk. However, FAIMs will make it harder to manage such risks through institution-by-institution action alone.[20] Moreover, such information registers might not fully capture dependencies further down the supply chain, including dependencies on essential underlying technology and open source components.

Given the shared dependencies across the financial sector, mitigation efforts undertaken by individual institutions are unlikely to be enough where vulnerabilities affect critical third-party providers, shared technological ecosystems or widely used open source components. Effective response and risk mitigation therefore require a coordinated approach involving all affected parties, including AI providers, software providers, security firms, open source maintainers, financial institutions and authorities at both national and EU level.

3.3 Geopolitical dimension

The geopolitical dimension of FAIM development has already been discussed from an industrial perspective. Another important consideration is that a systematic lack of defensive cyber capabilities in the European Union could also erode confidence in the EU financial system globally, increasing funding costs and prompting business to flow towards other jurisdictions. While current export restrictions on FAIMs appear to be limited in scope, they might plausibly be extended in the future. This should be considered in light of existing global restrictions on advanced semiconductors and rare earth minerals, along with controls over critical cloud infrastructure, which are not fundamentally different from restricting the use of FAIMs. Furthermore, FAIMs could come to be permanently classified as dual use technology by other jurisdictions and therefore become subject to rigorous export controls similar to those applied to other dual or military use technologies, potentially extending even to industry research and academia.

Geographical concentration among AI providers might also increase dependence on vendors subject to the extraterritorial legislation of third

[20] As some financial institutions are exempt from maintaining such a register, it cannot be fully relied on to provide a comprehensive system-wide perspective.

countries and forced to act under their local governments. This could include restricting access to the most advanced features, thus giving non-EU financial institutions a significant competitive advantage and ultimately leaving the EU exposed to strategic dependency and geopolitical risk. These exposures are more pronounced in the EU than in other large jurisdictions and raise broader questions about digital autonomy, which go beyond the scope of this note.

The AI Act extends to all providers of general purpose AI (GPAI) models regardless of where the provider is established once the model is placed on the EU market or its output is used within the EU, encompassing all models relevant to this note. The AI Act also imposes further requirements on GPAI models classified as posing systemic risk owing to their capabilities, whether presumed or evaluated. Models may also be subject to additional obligations following designation by the European Commission. Therefore, the interpretation and enforcement of the AI Act offer possible avenues to address some of the concerns raised in this note.

4 Conclusions

Systemic cyber incidents could propagate across the broader financial system through payment systems, clearing and settlement, or other operational bottlenecks, and could originate both within and outside the EU. By increasing the speed at which such attacks can be carried out, the use of FAIMs has the potential to amplify their impact. FAIMs also change the likelihood that an initial attack will affect a large number of institutions simultaneously, or severely disrupt a key financial institution, shocking the system.

FAIM-powered attacks could also be used to weaken the capabilities of financial institutions over time. A private financial institution might be forced to continuously increase its cybersecurity and operational resilience budget by investing in advanced defensive capabilities and specialised personnel to investigate, contain and remediate cyber incidents. This could have the effect of diverting resources away from core business, making the institution less profitable, undermining shareholder confidence and ultimately weakening overall economic resilience. Similarly, a persistent increase in cyber risk, coupled with the materialisation of such risks, could impair financial institutions and make them more exposed to other structural vulnerabilities. FAIM development could also have adverse implications for relevant authorities, forcing them to adapt their toolkit, build expertise and design approaches to keep pace with a rapidly evolving technological frontier, including by calibrating their expectations, stress testing and preparedness measures.

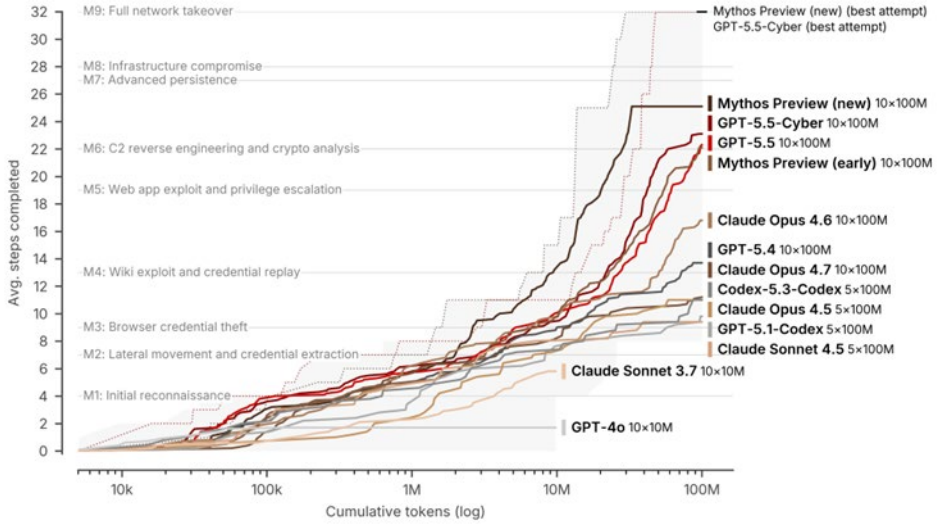
FAIMs may have a direct adverse effect on public confidence in the financial system. Specifically, any large-scale or correlated disruption affecting critical or important functions, especially in not readily substitutable financial market infrastructures, could erode the general public's trust in financial institutions. Even where such disruptions are temporary, repeated or concurrent events may create a perception of fragility, amplifying uncertainty and triggering irrational responses, including the transfer of funds to perceived safer alternatives, such as in other jurisdictions. As FAIMs capabilities continue to advance, the likelihood of confidence-related risks materialising may increase, potentially weakening overall stability and contributing to greater financial volatility.

Appendix A – UK AI Security Institute figure on model performance

The figure shows the average number of steps completed on “The Last Ones” (a 32-step simulated corporate network attack) as a function of total tokens spent. Each line represents a different model, with the shaded region showing the min-max range across all runs at each token budget. Grey horizontal lines indicate significant milestones in the attack chain.

Figure A1

Completed steps on “The Last Ones” per spent tokens



Source: UK AI Security Institute (2026), “How fast is autonomous AI cyber capability advancing?”, 13 May.

Appendix B – Illustrative hypothetical scenario narratives

The three scenarios are presented in order of severity, with efforts made to minimise the overlap between them.

Gradual erosion of trust in non-critical institutions

Threat actors: organised criminal groups (OCGs), as well as state-affiliated threat actors seeking to profit from cybercrime.

Motivation: monetisation of successful attacks, primarily by stealing assets, initiating large transfers, deploying ransomware, or a combination thereof.

Targets: small and medium-sized banks, non-bank financial institutions (NBFIs), and crypto-asset and tokenisation institutions.

In an evolving cyber landscape affected by recent FAIM advancements, OCGs and other financially motivated threat actors may gain access to FAIMs in pursuit of financial gain. Threat intelligence indicates that a growing share of recent attacks show signs of FAIM use. The primary targets are financial institutions with a relatively weak cybersecurity posture, combined with either direct access to highly liquid assets that can easily be stolen or transferred, or the ability to initiate large transfers of capital within the banking system. More precisely, attacks have tended to target smaller banks, NBFIs and smaller institutions in the digital asset ecosystem holding cryptocurrency or tokenised assets.

At the institutional level, there is no sanctioned access to FAIMs within the EU. However, limited-capability FAIMs have been appearing for sale on the dark web and through other illicit channels, although it remains unclear whether they are genuine FAIMs or merely rebranded general purpose AI models. There has been no coordinated EU-level response, although some private institutions have tested unsanctioned models with limited success. Moreover, several financial authorities have published guidance on how to prepare for the proliferation of FAIMs.

So far, the attacks have not had any systemic impact, as they have remained idiosyncratic and local, carried out by loosely affiliated threat actors against isolated targets. NBFIs have been able to absorb most of the damage, although losses have exceeded risk tolerance thresholds in some cases. Meanwhile, banks have proved more resilient, sustaining only minor losses in most cases and with only one bank being put into resolution. However, several individual crypto-asset service providers have suffered catastrophic damage, with large volumes of data either stolen or erased, in some cases leading to bankruptcy. Despite this, the broader ecosystem remains operational.

The initial impact of FAIM-enabled attacks has been concentrated in the crypto-asset ecosystem, resulting in significant financial losses and a loss of confidence in the EU digital asset ecosystem. Deteriorating capital buffers and relatively limited financial damage to NBFIs have led to reduced profitability and investor dissatisfaction. Meanwhile, customers of smaller banks have transferred their deposits and portfolios to larger institutions, fearing that otherwise they might be affected by a cyber incident. The impact on smaller banks and NBFIs has resulted in a partial loss of confidence, triggering a flow of capital from smaller EU banks and NBFIs to larger EU banks, and from EU NBFIs to NBFIs located in other jurisdictions. In the medium term, rating agencies structurally downgrade the ratings of smaller EU banks and NBFIs, citing increased cyber stress within the system, further driving the outflow of capital and pushing up risk premia.

Strategic intelligence gathering by banks and authorities

Threat actors: highly capable nation states and affiliated intelligence services.

Motivation: espionage and long-term strategic advantage.

Targets: large EU banks and financial authorities (including central banks and competent authorities).

Rather than becoming widely proliferated, FAIMs remain concentrated in a relatively small number of highly capable nation states. Military contracts and close cooperation with governments have provided intelligence agencies and affiliated threat actors in these countries with powerful espionage and cyber tools. These tools are used primarily against EU financial authorities and large EU banks, enabling intelligence agencies to establish a long-term presence within the EU financial system for strategic intelligence gathering purposes. Avoiding detection and long-term persistence are primary objectives. Access to FAIMs remains largely restricted, with the most capable models accessible only through foreign government contracts with AI providers.

Following the initial waves of attacks, threat actors gradually and systematically establish persistent access across the EU financial sector. Within three months, they are well established within most large banks and financial authorities, having secured multiple footholds across their systems. During this time, the EU takes isolated actions in an effort to strengthen the operational resilience and cybersecurity of the financial system, but is unsuccessful in securing access to FAIMs for financial authorities. This severely impairs their ability to detect, contain and eradicate FAIM-driven attacks on their infrastructures, leaving the sector response fragmented and ineffective.

In the short term, the operational effects of the attacks are only minor and the financial system continues to operate normally. Larger banks are able to detect anomalies in their ICT infrastructure, but are unsure if they can eradicate them and

generally attribute them to system noise. The ECB and several national central banks detect an increase in sophisticated attacks but remain convinced that they have not been compromised.

The medium and long-term effects of sustained, industrial-scale espionage place EU banks at a structural disadvantage. Sensitive information, including strategic, operational and commercial information, is continuously transferred from EU institutions to non-EU institutions, which use it to gain forward-looking insights and secure a strategic advantage. Over time, market distortion accumulates, placing EU institutions at a systemic disadvantage. Moreover, this information asymmetry allows non-EU actors to anticipate and pre-empt strategic initiatives, gaining advantages in both financial and non-financial markets. EU institutions face systematically disadvantaged competitive conditions through margin compression and weakened positioning in tenders and negotiations. Over time, this dynamic leads to a reallocation of market share towards non-EU actors, further strengthening their competitive advantage.

Furthermore, the ability of threat actors to obtain sensitive data on individuals and customers risks further eroding trust within the EU and increasing anxiety over data privacy and security.

The risk of public exposure is not one-sided and may also affect the originating jurisdictions themselves.

While the financial system remains fully operational, the main consequences would be a loss of confidentiality, a decline in trust and an erosion of the competitive advantage for private institutions and policymakers alike, rather than immediate systemic disruption.

Coordinated disruption of core infrastructure

Threat actors: second-tier cyber-capability nation states and affiliated threat actors.

Motivation: disruption of the financial system and erosion of public confidence.

Targets: financial market infrastructures (FMIs), including payment systems, clearing and settlement systems, and systemically important financial institutions.

In this scenario, a highly capable nation state obtains access to a fully functioning version of a specific FAIM. Threat intelligence indicates that the model is likely to leak into broader circulation within a matter of months, but incorrectly concludes that the threat actor will use the model primarily for military purposes, targeting the defence industry, the energy sector, and water and sanitation plants.

The threat actors instead immediately begin leveraging the model to identify vulnerabilities and establish footholds within EU FMIs, all with the goal of launching a coordinated and synchronised disruption of core financial services. In parallel, preparations are made for a large-scale misinformation, disinformation and

malinformation (MDM) campaign intended to amplify public distress and erode confidence in the financial system.

While some FMIs are able to detect certain intrusion attempts, they are unable to detect and respond to the more sophisticated attacks. At the same time, EU institutions lack access to comparable FAIM capabilities and have not developed a coordinated, system-wide solution for sharing indicators of FAIM compromise.

Given the current state of operational resilience, FMI contingency solutions are unable to absorb the impact of such a large-scale, synchronised attack. Although intelligence sources report the possibility of an imminent MDM campaign, they are unable to substantiate the nature, timing and scale of the threat.

When the threat actor can no longer conceal its use of the model and fears that it may leak into the public domain, it brings forward its plans and initiates a coordinated attack targeting core FMIs, hoping that whatever foothold it had acquired will suffice. The attack is accompanied by the activation of the MDM campaign.

As the attack on the FMIs plays out, in conjunction with the MDM campaign, the threat actor leaks the model through selected channels, ensuring that OCGs and hacktivists also gain access to FAIMs in the hope of further destabilising the EU financial sector. This has the effect of allowing more opportunistic threat actors to take part in the attacks, causing further disruption across the EU financial system.

The combination of targeted disruption of FMIs and widespread opportunistic attacks results in significant disruption to key financial services and systems. The impact is systemic, causing severe short and long-term damage to the EU financial sector.

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.esrb.europa.eu

The cut-off date for the data included in this report was 01 June 2026

For specific terminology please refer to the ESRB glossary (available in English only).

PDF ISBN 978-92-9472-471-7, doi:10.2849/4575934, DT-01-26-010-EN-N