Box 5.3

FINANCIAL STABILITY IMPLICATIONS OF TECHNOLOGICAL DISRUPTION IN CYBER SECURITY

Rapid technological innovation is reshaping the cyber risk profile facing the financial system. Advances in artificial intelligence (AI) and other technologies, such as quantum computing, could accelerate the development of tools capable of compromising the security of information systems and result in a qualitative leap in their capabilities. This would require at least a partial reassessment of existing cyber security strategies.

This box reviews the most recent developments in this area, identifies their implications for financial stability and assesses the ability of the banking system and its regulatory and supervisory framework to adapt to and contain the operational risks.

Using AI to exploit software vulnerabilities

Much of the recent attention surrounding this issue is rooted in the emergence of Claude Mythos, an AI model developed by the US company Anthropic and designed for advanced reasoning and programming tasks. According to its developer, the model has demonstrated exceptional capabilities in identifying and chaining together software exploits.[1]

Based on information published by Anthropic on its technical website and in the documentation for Claude Mythos, the model has reportedly been capable of identifying a large number of critical zero-day vulnerabilities across a range of software components, some of which had existed undetected for decades.[2]

The novelty of Claude Mythos in terms of detecting vulnerabilities lies in its ability to do so automatically and then combine them into exploit chains, drastically reducing the time between identification and potential malicious use. Broadly speaking, such vulnerabilities could be exploited to gain unauthorised control of systems, access sensitive information and compromise the integrity and availability of critical services.

Vulnerabilities have been identified in several different software components, including operating systems, browsers, multimedia applications and libraries. While not necessarily the case, library vulnerabilities could potentially be exploited to forge certificates or decrypt private communications.

It should be stressed, however, that at this stage the available information is limited to Anthropic's own announcements; the capabilities of Claude Mythos have not yet been fully and independently validated and the disclosure may also function, in part, as commercial positioning in a highly competitive market for advanced AI models.

Under adverse scenarios, such technological developments could have negative systemic effects. Faster and larger-scale exploitation of software vulnerabilities could lead to more synchronised and extensive waves of cyber attacks that affect technologies shared by a large part of the financial sector and other sectors. In particular, exploits leveraging vulnerabilities in critical components of payment systems could disrupt or temporarily curtail economic transactions. Given the widespread use of certain technologies (such as operating systems) and the reliance of a great many firms on a small number of critical software and hardware providers, the scenario in which AI attains a high level of capability for use in cyber attacks would heighten the systemic dimension of operational risk.

The case of Claude Mythos should be understood within a broader trend in AI-based model development. Even if this specific model were ultimately not found to be markedly superior to existing alternatives, it would nevertheless represent another step in the steady enhancement of the toolkit available not only to legitimate users, but also to potentially malicious actors. In any case, the likely emergence of multiple tools with advanced capabilities will facilitate their widespread adoption and further extend the systemic relevance of these technological developments.

Against this backdrop, Anthropic has launched Project Glasswing, providing invite-only access to Claude Mythos to a small group of US technology, cyber security, financial and open-source software firms. The aim is to enable the

1 Chaining software exploits means sequentially combining multiple, distinct security flaws to mount an attack with a much greater impact than exploiting any single vulnerability. For example, one exploit may be used to gain initial access to a system, a second to escalate access privileges and a third to exfiltrate data.

2 A zero-day vulnerability is a security flaw in software or hardware that is unknown to the developer and for which no patch or mitigation is available at the time it is exploited by an attacker. The term reflects the fact that developers have "zero days" to address the vulnerability before it can be used for malicious purposes.

early development of defensive strategies ahead of any broader commercial release.

While this initiative may help identify and mitigate risks associated with the technology, effective containment of its systemic dimension will require close international coordination. In a global economy with significant cross-border interconnections between firms and institutions, failure to mitigate vulnerabilities across large parts of the network could facilitate attacks that propagate to the global system as a whole. The Financial Stability Board[3] could, in particular, serve as one of the forums to facilitate such international coordination.

Given the systemic and global nature of these technologies, it is necessary for firms outside the United States to have access to Project Glasswing or similar initiatives, alongside regular exchanges between the public and private sectors. Where risks are potentially systemic, there must be global mechanisms for analysing and testing new technologies and sharing findings that avoid asymmetries.

It is also important for Europe to develop robust capabilities in these technologies, advancing digital sovereignty, strengthening its own cyber defence capacity and ensuring it plays an active role in designing and implementing global solutions to these risks.

Crucially, the potential use of these technologies to strengthen defence systems and reduce reaction times to vulnerabilities will require banks to adopt more agile decision-making models and creates new uncertainties regarding the technical quality of remediation solutions. Overall, the balance between AI-driven improvements in cyber offence and defence capabilities remains uncertain. Figure 1 summarises the main elements of the publicly available information about Claude Mythos and its implications for operational risk.

Quantum computing and cryptography

Quantum computing may ultimately have even more far-reaching implications for cyber security than AI, although it is at an earlier stage of development and its future path is subject to greater uncertainty. While its practical impact is likely to remain limited in the short term, advances in quantum computing pose significant risks to existing cryptographic standards, which underpin the confidentiality, integrity and authenticity of financial transactions. A future transition to post-quantum cryptography would entail substantial technical and organisational efforts and therefore requires advance planning.

Quantum computing is a computational paradigm that makes use of the principles of quantum mechanics (a field

Figure 1
The Claude Mythos AI model

	What is it?	Why does it matter?
Claude Mythos	• An AI model with advanced capabilities in identifying and exploiting software vulnerabilities • Developed by the US company Anthropic	• It could reshape the threat model, accelerating and scaling up the identification and exploitation of vulnerabilities in information systems
Main technological risk	• Increase in the speed, scale and automation of attacks on information systems • Its capabilities are currently uncertain, pending independent and wide-ranging verification	• Current defences may be overwhelmed by the speed and new types of cyber attacks • It could generate more synchronised waves of vulnerability exploits that are difficult to handle using existing systems
Project Glasswing	• It offers limited access to Claude Mythos to several US firms to develop defence strategies before a potentially wider commercial release	• It gives participating firms a head start to develop defence strategies to address cyber attacks leveraging this technology
Systemic implications	• Increase in cyber incidents in all areas of the economy. • The capabilities of Claude Mythos can be replicated and enhanced by other models	• Increase in operational risks across all economic sectors

SOURCE: Banco de España.

3 The Financial Stability Board is an international body established in 2009 that monitors the global financial system and recommends policies to bolster its resilience. It contributes to the global monitoring of vulnerabilities and risks to financial stability and promotes the consistent implementation of standards and coordination of financial stability policy.

of physics) to process information, allowing certain problems to be solved far more quickly than would be the case with classical computers.[4] Much of the security of current cryptography relies on the complexity of specific mathematical problems that are extremely hard to solve for conventional computers, but significantly easier for quantum ones. That said, quantum technologies still face major challenges to large-scale practical deployment, including error correction, decoherence, high energy costs, extreme physical operating conditions and the absence of standardised solutions that can be applied at scale.

Despite these challenges, two articles published in March this year[5] highlighted the risk that the time frame for the emergence of machines capable of breaking parts of current cryptography may be shorter than previously believed. They suggest that the quantum resources required to attack certain public-key schemes could be substantially lower than previously estimated.[6] While the risk is not imminent as yet, it has become more credible and has moved closer from a prudential perspective, underscoring the need for clear understanding of cryptographic inventories, sound governance arrangements and close engagement with service providers. Figure 2 provides an overview of differences in the assessment of operational risks associated with AI and quantum computing.

Financial stability impacts beyond direct operational risk

From a financial stability perspective, the effect of such technological threats on confidence in the financial system is also relevant. At bank level, there is already some evidence that cyber attacks can trigger fund outflows (sudden or gradual) towards banks or investments perceived as more resilient, even when that perception is not objectively warranted.[7]

However, for such events to escalate into a systemic episode of volatile fund flows and, therefore, liquidity

Figure 2
Comparison of technological development and the risk implications of AI and quantum computing

	AI	Quantum computing
Potential impact on cyber security	It could accelerate and automate vulnerability identification and exploitation. It could cut down the delay between discovery and attack, but also improve defences	Disruptive risk for current cryptography. Uncertain future impact, concentrated in systems that have not migrated to post-quantum standards
Maturity	High in general applications. Potentially rapid spread of advanced models with emerging cyber security capabilities	Low/medium. Significant advances in hardware and error correction, but lacking widespread applications outside experimental environments
Additional development periods	Short-term (one to three years). Observable and scalable operational uses	Medium to long term. Potentially high impact, but with an uncertain time frame
Probability of materialising	High. Low barriers to entry, repurposing of existing infrastructures and rapid adoption by legitimate and malicious actors	Medium to low. It depends on technical milestones that are not yet fully established and on sustained investment over time

SOURCE: Banco de España.

4 Advances in quantum computing are part of a broader set of developments in quantum technologies. Other prominent areas of research include quantum communications and cryptography (which could potentially enhance the security of information transmission and encryption), as well as quantum sensing and metrology (which could enable more precise measurement of a wide range of physical phenomena than is possible with current technologies).

5 See Babbush et al. (2026), Securing Elliptic Curve Cryptocurrencies against Quantum Vulnerabilities: Resource Estimates and Mitigations, arXiv:2603.28846 and Cain et al. (2026), Shor's algorithm is possible with as few as 10,000 reconfigurable atomic qubits, arXiv:2603.28627

6 Public-key (or asymmetric) cryptography uses a pair of mathematically related keys: a public key (freely shared for encryption and verification) and a private key (kept secret for decryption and signing).

7 See A. Kotidis and S. Schreft. (2025). "The Propagation of Cyberattacks through the Financial System: Evidence from an Actual Event". Journal of Finance, 80, pp. 3313-3358; and F. Gogolin, I. Lim and F. Vallascas. (2026). "Cyberattacks on Small Banks and the Impact on Local Banking Markets". Journal of Money, Credit and Banking, forthcoming. The first study shows that attacks on banks' information and communication technology third-party providers can disrupt payment systems and give rise to liquidity stress, while the second reports that cyber attacks are associated with deposit outflows.

stress, misuse of these AI models would need to enable simultaneous, severe and prolonged attacks on multiple banks (or on providers widely used in the financial sector) to undermine confidence in a systemically important number of intermediaries or financial instruments. Such scenarios differ from isolated cyber incidents, which are unlikely to have significant implications for financial stability.[8]

Another channel through which these technologies could affect financial stability is their medium-term impact on the profitability of banks (and other intermediaries), should a scenario of severe and recurrent cyber incidents materialise. First, banks may be obliged to make larger and more frequent investments in cyber security to address these new threats. Second, funding costs could rise for certain products or parties owing to the aforementioned reallocation between banks or financial instruments during episodes of loss of confidence.

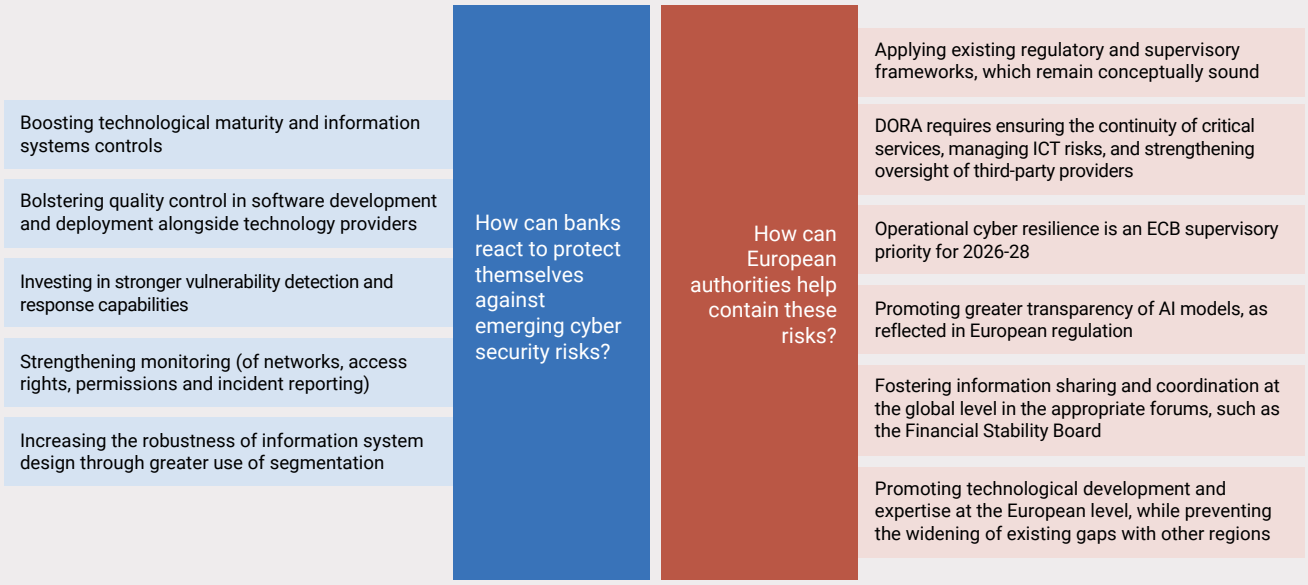
On the asset side, a sustained increase in cyber attacks could impair the solvency of the more exposed corporate sectors, with potential repercussions for banks' credit and market risk losses.[9] In short, by affecting profitability, cyber attacks can weaken the banking system's resilience to other types of shocks.

Response capacity of the banking system and the regulatory and supervisory framework

A higher initial exposure to these risks could prompt additional strengthening of banks' cyber resilience and adjustments to the regulatory and supervisory framework, thereby helping to preserve public confidence in the banking system as a whole.[10]

The risk scenarios associated with these technologies underline the need for technological maturity and robust

Figure 3
Capacity of banks and the supervisory and regulatory system to respond to emerging risks



SOURCE: Banco de España.

8 For a broader discussion of the potential for cyber attacks to develop into a financial stability crisis, see Section 5.4 of the Financial Stability Report. Spring 2025.

9 See R. Jamilow, H. Rey, and A. Tahoun. (2025). "The Anatomy of Cyber Risk". NBER Working Paper 28906. P. Akey, S. Lewellen, I. Liskovich, C. Schiller. (2026). "Hacking Corporate Reputations". The Review of Finance (forthcoming). These studies show that greater corporate exposure to cyber attacks is associated with lower profitability.

10 In particular, the adoption of supervisory measures aimed at ensuring minimum standards could encourage banks to invest in cyber security. See T. Ahnert, M. Brolley, D. Cimon and R. Riordan (2024), "Cyber Risk and Security Investment"; S. Kokas, F. Vallascas and N. Rowe (2026), "Digitalization in Banking: Evidence from the New York Cybersecurity Regulation"; and K. Anand, C. Duley and P. Gai. (2026), "Cybersecurity and Financial Stability". The Review of Finance (forthcoming).

controls in banks' information systems. As a first line of defence, banks could bolster quality control in the development and deployment of new software solutions, working closely with their technology providers (e.g. through more intensive testing prior to deployment) to limit exploitable vulnerabilities.

In addition, investments should be made in enhancing responsiveness to vulnerabilities, shortening remediation times and improving detection capacity. Better monitoring of network activity, permission and access control systems and cyber incident reporting would all be useful for this purpose.[11] As already noted, advances in AI technology can also be harnessed for this purpose, for example, to harden information systems by identifying vulnerabilities when they are in the design phase.

In the face of these technological challenges, existing regulatory and supervisory frameworks remain conceptually sound and already contain the core controls needed to mitigate the risk arising from these developments. For instance, the Digital Operational Resilience Act (DORA), in force since January 2025, sets out key requirements for security policies, regular testing and managing technological and third-party risks, such as information and communication technology (ICT) third-party service providers.

On the supervisory side, ECB Banking Supervision has placed operational cyber resilience at the centre of its supervisory priorities for the 2026-28 cycle[12] and has expanded direct oversight to critical ICT providers. This framework is complemented by advanced testing exercises, such as threat-led penetration testing, which assess the ability of the largest banks to detect, respond to and recover from sophisticated cyber attacks, thereby strengthening the operational resilience of the financial system as a whole.

Notwithstanding these strengths, existing frameworks were designed under different assumptions regarding the speed, automation and scale of attackers. The challenge is not to create entirely new regimes, but to update assumptions, scenarios and practices, enhancing testing approaches, response plans and governance mechanisms that allow for agile decision-making in high-uncertainty environments. In threat-led penetration testing exercises, for example, it is essential to ensure that testing providers have sufficient technical capabilities to emulate the most sophisticated attacks, including those using models such as Claude Mythos.

Paradoxically, despite the rapid evolution of these AI models, transparency standards among leading providers of these technologies, such as large language models (LLMs),[13] are not improving. The Foundation Model Transparency Index, compiled by researchers from Stanford, Princeton and MIT, shows that LLMs comply, on average, with fewer than half of the transparency requirements. The overall score stands at 41% and, notably, represents a 17-point slide relative to 2024, with transparency effectively falling back to 2023 levels.[14] This clearly suboptimal outcome highlights the need for progress in this area. Transparency is a cornerstone of the European Union's AI Act and delays in its implementation are not helping to raise standards.

Conclusions

Alarmist interpretations of the potential impact of cyber risks on financial stability should be avoided. There is no evidence of an immediate systemic threat stemming from AI models such as Mythos or from advances in quantum computing. Experience nevertheless shows that technological advances tend to proliferate rapidly. This creates a limited window for preparedness against cyber risks, during which it is essential to enhance technical capabilities, improve coordination among banks, supervisors and providers, and invest in specialised talent. International coordination is particularly critical during this preparatory stage to bolster global resilience and prevent attacks that initially affect less-prepared parties from spreading through interconnections.

Ultimately, in an environment of accelerated technological disruption, the financial system's cyber resilience will depend above all on the speed of detection, remediation and coordination, as well as on the robustness of governance arrangements and the anticipation of emerging risks.

11 Cyber incidents are events that affect the confidentiality, integrity, availability or authenticity of information systems as a consequence of technical failure, human error or malicious activity.

12 See, for example, ECB (2026) "Upgrading banks' capacity to deal with digital risks", Contribution by Anneli Tuominen, Member of the Supervisory Board of the ECB, for Eurofi Magazine.

13 LLMs are trained on large text corpora, which enables them to understand and generate natural language.

14 See, for reference, the Foundational Model Transparency Index.