INVESTIGACIONES DE FUNCAS
40/2026

ARTIFICIAL INTELLIGENCE AND FINANCIAL STABILITY: BANKING RISKS IN THE AGE OF FRONTIER AI — A COMPARATIVE ASSESSMENT OF SPAIN, THE EUROPEAN UNION, AND THE UNITED STATES

Francisco Rodríguez Fernández

June 2026

funcas

Edita: Funcas
Caballero de Gracia, 28, 28013 - Madrid

ISSN: 3045-8994

ARTIFICIAL INTELLIGENCE AND FINANCIAL STABILITY: BANKING RISKS IN THE AGE OF FRONTIER AI — A COMPARATIVE ASSESSMENT OF SPAIN, THE EUROPEAN UNION, AND THE UNITED STATES

Francisco Rodríguez Fernández

University of Granada and Funcas

Abstract

This paper takes the May 2026 Anthropic–FSB briefing on the Claude Mythos Preview model as the empirical anchor for a comparative assessment of how Spain, the European Union, and the United States are positioned to manage AI-related banking risks. We argue three theses. First, frontier AI has crossed a capability threshold — evidenced by Mythos's 83.1% pass rate on the CyberGym vulnerability-reproduction benchmark — that transforms cyber risk from an idiosyncratic operational hazard into a correlated, scalable, and partly endogenous systemic risk. Second, the existing regulatory architecture across the three jurisdictions is necessary but insufficient: the U.S. recently excluded generative and agentic AI from its revised model risk guidance (SR 26-2); the EU has the most prescriptive horizontal framework (AI Act, DORA) but is in early implementation; and Spain has front-run EU peers via AESIA while operating with materially smaller technical resources than its peer central banks. Third, dependence on a few frontier model providers and cloud hyperscalers — with the top three providers accounting for 44% of third-party AI models used by UK financial firms — converts an industrial-organization problem into a financial-stability one. We develop a six-dimension risk taxonomy, present four scenario analyses calibrated to real-world benchmarks (SVB 2023, the 2010 flash crash, the August 2024 yen carry unwind, and the July 2024 CrowdStrike outage), and provide a Spain-specific deep dive analysis. We conclude with seven policy recommendations across supervisors, macroprudential authorities, industry, and frontier AI labs.

Keywords: Artificial intelligence, financial stability, banking supervision, cyber risk, frontier models, systemic risk, EU AI Act, DORA, Spain.

JEL Classification: G18, G21, G28, O33, K23.

EXECUTIVE SUMMARY

On May 18, 2026, the Financial Times reported that Anthropic had agreed to brief the Financial Stability Board (FSB) on cyber vulnerabilities in the global financial system identified by its Claude Mythos Preview model, at the request of Bank of England Governor and FSB Chair Andrew Bailey (Financial Times, 2026; Reuters, 2026). The briefing — preceded by a April 7–8, 2026 closed-door meeting between U.S. Treasury Secretary Scott Bessent, Federal Reserve Chair Jerome Powell, and major U.S. bank CEOs (CNBC, 2026; Sullivan & Cromwell, 2026) — marked the first time a frontier AI laboratory was treated as a de facto systemic-relevance actor by global financial regulators. Bailey, speaking at Columbia University on April 14, 2026, warned that Mythos could "crack the whole cyber risk world open" (Bank of England, 2026; Reuters, 2026).

The Mythos event is a very relevant feature, but the underlying phenomenon is structural. Roughly 90% of significant euro-area banks already use AI (European Central Bank, 2026), 75% of UK financial firms have adopted some form of AI (Breeden, 2024), and the top three external AI providers now account for 44% of third-party AI models used by UK financial institutions, up from 18% in 2022 (Breeden, 2024). The Financial Stability Board (2024) has formally identified four AI-related vulnerabilities — third-party dependencies, market correlations, cyber, and model risk — that materially affect financial stability. The European Systemic Risk Board's Advisory Scientific Committee (2025) concluded that five AI features — concentration and entry barriers, model uniformity, monitoring challenges, overreliance, and speed — could significantly amplify systemic risk.

This paper argues three theses. First, frontier AI has crossed a capability threshold that transforms cyber risk from an idiosyncratic operational hazard into a correlated, scalable, and partly endogenous systemic risk for the banking system, requiring macroprudential — not just microprudential — treatment. Second, the existing regulatory architecture across the U.S., EU, and Spain is necessary but insufficient: the U.S. relies on principles-based supervisory letters and recently excluded generative and agentic AI from its revised model risk guidance (Federal Reserve Board et al., 2026); the EU has the most prescriptive horizontal framework (EU AI Act, DORA) but is still in early implementation; and Spain has front-run EU peers by creating the first national AI supervisory agency (AESIA) but operates with substantially smaller resources than central banks. Third, dependence on a small handful of frontier model providers and cloud hyperscalers — AWS at approx. 30%, Azure at approx. 21%, Google Cloud at approx. 13% of global cloud (Synergy Research Group, 2025), with NVIDIA at approx. 92% of data-center GPUs (IoT Analytics, 2025) — converts an industrial-organization problem into a financial-stability one, with concentration risk now mirroring the "too-big-to-fail" externalities of the largest financial intermediaries.

Four scenario analyses, calibrated to real-world benchmarks (SVB March 2023, May 6, 2010 flash crash, August 5, 2024 yen carry-trade unwind, July 19, 2024 CrowdStrike outage), suggest that an AI-amplified deposit-run event at a mid-sized G-SIB could plausibly compress the SVB outflow timeline of $42 billion in one day (Federal Reserve Board, 2023) further still, while a Mythos-class autonomous-attack scenario could shrink the median industry "time-to-exploit" — for which the historical median was months — to hours or minutes (Anthropic, 2026a; Cisco's Anthony Grieco, quoted in Anthropic, 2026a). Spanish banks, while well-capitalized (sector CET1 of 13.8% as of June 2025; Banco de España, 2025a), face elevated relative exposure due to high domestic concentration (CR3 of approx. 72%; World Bank, 2023), Santander's G-SIB status, and substantial Latin American and Turkish footprints.

The paper concludes with policy recommendations across four constituencies: supervisors, macroprudential authorities, industry, and frontier AI labs. Most urgently, it argues that the FSB's planned Sound Practices consultation (FSB, 2026) should formalize three principles: (i) designation of critical AI providers analogous to DORA's CTPP regime; (ii) mandatory pre-deployment red-teaming by independent national institutes (UK AISI, U.S. CAISI) for any frontier model offered to systemic banks; and (iii) a coordinated "responsible disclosure" protocol for AI-discovered vulnerabilities in financial infrastructure, modeled on the bespoke Anthropic–FSB channel that the Mythos affair has improvised.

1. INTRODUCTION: THE MYTHOS MOMENT AND THE AI–FINANCE NEXUS

1.1. An unprecedented regulatory engagement

The April–May 2026 sequence of events is without precedent in the post-1944 history of international financial governance. Anthropic announced Claude Mythos Preview on April 7, 2026, as part of "Project Glasswing,"describing it as a "general-purpose, unreleased frontier model that reveals a stark fact: AI models have reached a level of coding capability where they can surpass all but the most skilled humans at finding and exploiting software vulnerabilities" (Anthropic, 2026a). Within 24 hours, Treasury Secretary Bessent and Federal Reserve Chair Powell had convened the CEOs of Bank of America, Citi, Goldman Sachs, Morgan Stanley, and Wells Fargo, with Jamie Dimon of JPMorganChase invited but unable to attend (CNBC, 2026). Bailey named the model publicly seven days later (Bank of England, 2026), the Bank of England's Cross-Market Operational Resilience Group placed Mythos on its agenda within two weeks (Bloomberg, 2026a), and by mid-May 2026 the Australian Securities and Investments Commission, the European Central Bank, the Japanese Financial Services Agency, the Singapore MAS, and the South Korean FSC had all issued public statements or convened bank CEOs on the threat (Reuters, 2026; Nikkei Asia, 2026; NL Times, 2026).

The capabilities are extraordinary. Anthropic reports Mythos achieved an 83.1% pass rate on CyberGym's cybersecurity-vulnerability-reproduction benchmark, against 66.6% for Claude Opus 4.6 (Anthropic, 2026a). XBOW's independent evaluation found that Mythos cut false-negative bug discoveries by 42% versus Opus 4.6 in live web-vulnerability testing, and by 55% with source code provided (XBOW, 2026). The UK AI Security Institute reported that Mythos solved 22 of 32 steps in a "Last Ones" corporate-network attack simulation that takes human experts an estimated 20 hours, and that the doubling time of autonomous-cyber-task capability has compressed from approximately 8 months in November 2025 to roughly 4.7 months by February 2026 (UK AI Security Institute, 2026a, 2026b). Mozilla shipped fixes for 271 vulnerabilities in Firefox 150 found by Mythos in a single evaluation pass (The Next Web, 2026).

1.2. Why this is a financial-stability problem, not just a cybersecurity problem

The orthodox view in 2024 —articulated in the FSB's seminal Financial Stability Implications of Artificial Intelligence report (FSB, 2024)— was that AI introduces known operational and cyber vulnerabilities that can be addressed with adaptations of existing frameworks. The Mythos moment unsettles that view in three ways.

First, the speed of capability progression. CrowdStrike's CTO Elia Zaitsev observed that "the window between a vulnerability being discovered and being exploited by an adversary has collapsed —what once took months now happens in minutes with AI" (quoted in Anthropic, 2026a). When the time-to-exploit collapses below the time-to-patch, the equilibrium of cyber defense— a probabilistic race between defenders and attackers — shifts unfavorably for incumbents whose patch cycles are calibrated to the slower historical equilibrium.

Second, the concentration of frontier-AI capability. Approximately 40 organizations have access to Mythos Preview, with 12 named launch partners (AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, NVIDIA, Palo Alto Networks, and Anthropic itself) (Anthropic, 2026a). Anthropic agreed not to distribute the model more widely at the request of the White House (Financial Times, 2026; Yahoo Finance, 2026). The asymmetry is acute: every G-SIB faces the same residual threat universe, but only a fraction has access to the corresponding defensive capability. As the European Central Bank's Frank Elderson observed in May 2026, this is "not an excuse for inaction" — but it makes "even more critical that banks step up and act now" (NL Times, 2026).

Third, the correlated nature of the threat. The financial sector's reliance on a small number of common software stacks (Microsoft Windows, Linux, Chrome/Firefox, common Java frameworks, SWIFT, Fedwire, T2/TARGET2) means that a single class of Mythos-class exploit could simultaneously affect dozens or hundreds of institutions. The same correlation logic that drives systemic risk in financial exposures now applies to operational

and cyber exposures — and the regulatory toolkit for managing the former is far more developed than for the latter.

1.3. Scope and roadmap

This paper synthesizes regulatory, supervisory, academic, and industry evidence from 2018 to May 2026 to produce a comparative assessment of how Spain, the European Union, and the United States are positioned to manage AI-related financial-stability risk. Section 2 sets out the conceptual framework. Section 3 develops a taxonomy of AI-related banking risks. Section 4 discusses positive applications of AI as a counterweight. Section 5 compares the three jurisdictions' regulatory architectures. Section 6 builds a model of contagion channels. Section 7 presents four calibrated scenario analyses. Section 8 examines Spain's specific exposure profile. Section 9 sets out policy recommendations. Section 10 concludes.

2. CONCEPTUAL FRAMEWORK: HOW AI PROPAGATES RISK IN BANKING

2.1. The three transmission layers

AI propagates risk into the banking system through three nested layers. The innermost layer is firm-level: a bank's adoption of AI in credit scoring, fraud detection, trading, or customer service introduces model-specific risks of error, bias, and operational failure. This is the classical domain of microprudential model risk management — Federal Reserve SR 11-7 (Board of Governors of the Federal Reserve System & OCC, 2011) and its 2026 successor (SR 26-2; Federal Reserve Board, FDIC, & OCC, 2026), the EBA's machine-learning IRB framework (EBA, 2023), and the ECB's revised internal-models guide (ECB, 2025c).

The middle layer is sector-level: when many banks adopt similar AI systems trained on similar data with similar architectures, idiosyncratic firm-level risks become correlated. Algorithmic herding, common-source data poisoning, and synchronized procyclical responses to shocks transform model risk into market risk. The IMF (2024) and FSB (2024) identify this as the principal novel channel introduced by widespread AI adoption.

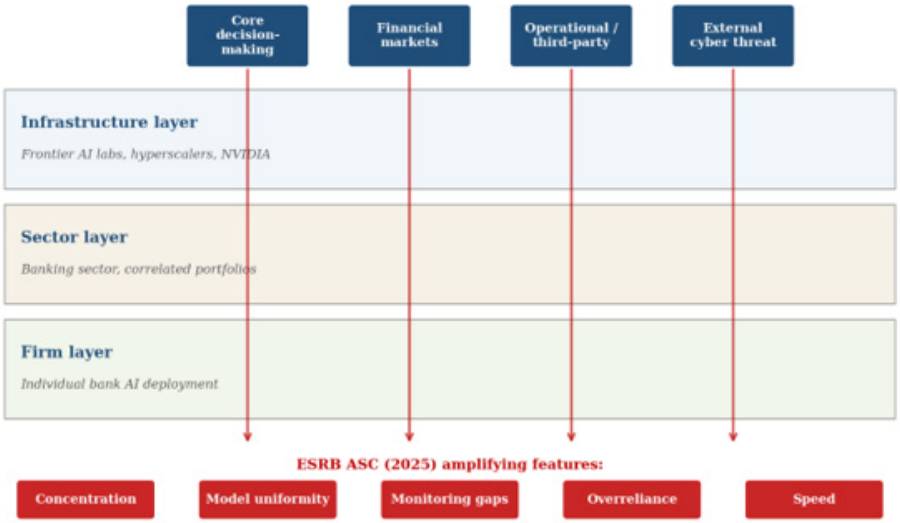
The outermost layer is infrastructure-level: dependencies on a few frontier model providers, cloud hyperscalers, and specialized hardware vendors (NVIDIA) create concentration risks that resemble the too-big-to-fail externalities of large financial intermediaries — but with critical regulatory perimeter problems. NVIDIA, Anthropic, OpenAI, AWS, Microsoft Azure, and Google Cloud are not within any banking regulator's direct supervisory perimeter, even as banking systems' operational continuity increasingly depends on them.

2.2. The five amplifying features of AI

The ESRB Advisory Scientific Committee (2025) provides the most rigorous typology of why AI is systemically distinctive. Of eleven features evaluated, five significantly amplify financial-system vulnerabilities: concentration and entry barriers, model uniformity, monitoring challenges, overreliance and excessive trust, and speed. Each map to one or more of the classical sources of systemic risk articulated in the macroprudential literature — liquidity mismatches, common exposures, interconnectedness, lack of substitutability, and leverage.

The Bank of England's Financial Policy Committee (Bank of England, 2025) identifies four channels through which AI-related risks transmit to financial stability: (i) greater use of AI in banks' core financial decision-making; (ii) greater use of AI in financial markets (correlated trading, herding); (iii) operational risks from AI service providers; and (iv) a changing external cyber-threat environment. These four channels can be mapped to the ESRB's five features and the FSB's four vulnerabilities to produce an integrated conceptual model, summarized in Figure 1.

Figure 1. INTEGRATED CONCEPTUAL MAP OF AI RISK TRANSMISSION TO BANKING STABILITY
(Three vertical layers x Four transmission channels x Five amplifying features)



Notes: (Conceptual diagram, not to scale).
Three vertical layers — firm, sector, infrastructure — intersect with four horizontal channels — core decision-making, markets, operational/third-party, external cyber threat. ESRB's five amplifying features (concentration, uniformity, monitoring challenges, overreliance, speed) sit as cross-cutting modifiers that magnify the magnitude and correlation of shocks transmitted through any combination of channel and layer.

2.3. Endogenous vs. exogenous risk

A distinctive feature of AI risk is that it sits uncomfortably between the standard categories of endogenous and exogenous risk in macroprudential theory. AI tools deployed by banks themselves contribute endogenously to potential procyclicality, herding, and correlated runs. AI tools deployed by adversaries —whether nation-state actors, cybercriminals, or activists— generate exogenous shocks that the financial system must absorb. Frontier models like Mythos sit on both sides of this divide simultaneously: the same capability that allows JPMorganChase or Microsoft to harden their systems also represents a latent threat if it reaches less-aligned actors. This duality, the dual-use problem, is the empirical heart of the Mythos affair and the most acute regulatory novelty of the moment.

3. TAXONOMY OF AI-RELATED BANKING RISKS

This section develops a six-dimension taxonomy. Table 1 summarizes the structure; the six subsections below provide depth.

3.1. Operational and cyber risk: The Mythos-type dual-use problem

The 2024 IMF Global Financial Stability Report estimated that "the once-in-a-decade cyber loss for a financial firm could reach $2.2 billion, up from approximately $300 million in 2017" (cited in IMF, 2026). The historical baseline is sobering: the SWIFT-Bangladesh Bank heist (February 2016) cost approximately $81 million of an attempted $951 million, with only $15 million recovered (BAE Systems, 2016; U.S. Department of Justice, 2018); the Capital One breach (July 2019) exposed 106 million customer records and produced an $80 million OCC civil money penalty plus a $190 million class-action settlement (OCC, 2020; Capital One Settlement Administration, 2022); the ICBC ransomware attack (November 8–9, 2023) disrupted settlement of over $9 billion in Treasury-

TABLE 1. SIX-DIMENSION TAXONOMY OF AI-RELATED BANKING RISKS

Dimension	Principal risk vectors	Key regulatory instruments	Empirical anchor
Operational & cyber	Frontier-model dual use; third-party concentration on AI providers; model risk management gaps	SR 26-2 (US); DORA Arts. 28–44 (EU); EU AI Act Art. 15; ECB cyber stress test framework	Mythos (April 2026); ICBC ransomware (Nov 2023); CrowdStrike outage (Jul 2024)
Systemic / macroprudential	Algorithmic herding; procyclicality; AI-driven flash events; AI-amplified deposit runs	ESRB ASC Report No. 16 (2025); IMF GFSR Oct 2024 Ch. 3; BoE FSiF Apr 2025	SVB run (Mar 2023); May 6, 2010 flash crash; Aug 5, 2024 yen carry unwind
Credit & market	AI credit scoring bias; algorithmic trading; deepfake-enabled fraud at scale	EU AI Act Annex III(5)(b); CFPB Circular 2023-03; MiFID II RTS 6	Arup HK$200m deepfake fraud (Jan 2024)
Governance, model risk, explainability	Black-box risk; hallucinations in customer-facing AI; model validation gaps	SR 26-2 (US); EBA ML-IRB Follow-up (2023); EU AI Act Arts. 13–15	ESMA finding: ChatGPT hallucinates approx. 19.5% of responses (2024)
Money laundering, fraud, conduct	Synthetic identity; voice cloning; deepfake CEO fraud; AI-generated disinformation	Treasury AI Cyber Report (Mar 2024); FBI IC3 reporting; FATF guidance	FBI IC3 $16.6B losses 2024; deepfakes +1,300% per Pindrop (2025)
Concentration / oligopoly	Dependence on Anthropic, OpenAI, Google DeepMind; NVIDIA dominance; AWS/ Azure/GCP triopoly	DORA CTPP regime; FSB 2024 recommendations	44% UK FS AI models from top 3 providers; NVIDIA approx. 92% data-center GPU

Source: Own elaboration from existing regulations.

backed assets (U.S. Department of the Treasury, 2024b); and the MOVEit zero-day campaign (May–June 2023) affected over 2,500 organizations and 66 million individuals, including numerous financial-sector victims (CISA & FBI, 2023; Emsisoft, 2023).

The introduction of frontier models fundamentally changes the threat surface. The U.S. Treasury's Managing Artificial Intelligence-Specific Cybersecurity Risks report (U.S. Department of the Treasury, 2024a) —issued under Executive Order 14110— identified four AI-specific vulnerabilities (data poisoning, data leakage during inference, evasion, and model extraction) and warned that "AI allows bad actors to impersonate individuals, such as employees and customers of financial institutions, in ways that were previously much more difficult." The March 2024 report also flagged the widening capability gap between large and small institutions as a sector-level vulnerability.

Model risk management has been the principal microprudential response. The U.S. Federal Reserve, FDIC, and OCC jointly issued SR 26-2 on April 17, 2026, superseding SR 11-7 (Federal Reserve Board, FDIC, & OCC, 2026). The revised guidance is roughly half the length of its 2011 predecessor and removes the annual revalidation requirement. Critically, generative AI and agentic AI are explicitly out of scope (SR 26-2, footnote 3). Vice Chair Bowman defended the choice on May 1, 2026: "The revised guidance now applies narrowly to traditional models and basic AI applications" (Bowman, 2026). The EU's approach is more prescriptive: high-risk AI systems under Article 6 of the AI Act must comply with Articles 9 (risk management), 10 (data governance), 11 (technical documentation), 13 (transparency), 14 (human oversight), and 15 (accuracy, robustness, cybersecurity) (Regulation (EU) 2024/1689). The ECB's revised Guide to Internal Models (ECB, 2025c) introduces, for the first time, ECB expectations on machine-learning models in IRB.

The third-party concentration angle is now arguably more important than the model-risk angle. The Bank of England's 2024 AI survey found that 44% of third-party AI models used by UK financial firms come from the top three providers, up from 18% in 2022 (Breeden, 2024). Anthropic's $30 billion Series G in February 2026 at a $380 billion post-money valuation (Anthropic, 2026b) and OpenAI's reported approx. $852 billion implied valuation (Bloomberg, 2026; cited in foreignpolicyjournal.com, 2026) make these providers more valuable than all but the largest banks, while operating entirely outside the prudential perimeter. Amazon's compute commitment to Anthropic totaling up to $25–$33 billion(Anthropic, 2025) and Microsoft's $13 billion stake in OpenAI (CNBC, 2025) further entangle frontier AI with cloud concentration.

3.2. Systemic and macroprudential risk

Three macroprudential channels deserve emphasis. First, algorithmic herding and procyclicality: the IMF (2024) reported that algorithmic trading already accounts for approx. 70% of U.S. equities and approx. 50% of futures volume. The Dutch energy-market authority found that "machine learning is implicitly or explicitly used in 80 to 100 percent of [trading firms'] algorithms" (cited in IMF, 2024). As Coalition Greenwich (2024) reported, 37% of U.S. equity buy-side order flow was executed via algorithms or smart order routers in 2023, up from 35% the year before. When models share architectures, data, or training methods, their responses to shocks correlate —which can be benign in calm markets and catastrophic in stressed ones. The IMF (2024) found that AI-powered ETFs experienced higher turnover than active or passive ETFs, with turnover spiking in March 2020— behavioral evidence of procyclicality.

Second, AI-amplified deposit runs. The Silicon Valley Bank failure of March 9–10, 2023 produced the empirical foundation for understanding social-media-driven runs. The Federal Reserve Board's (2023) review documents that "deposit outflows were over $40 billion on March 9, and management expected $100 billion more the next day" — implying total attempted outflows of approximately $142 billion against a deposit base of $166 billion (approximately 85%) compressed into 36 hours. Cookson et al. (2025) found that 6,528 "run" tweets about SVB were posted between March 8–13, 2023, and that banks in the top tercile of pre-existing Twitter exposure lost 6.6 percentage points more stock market value during the run period — an effect comparable in magnitude to one standard deviation of uninsured deposits. The implication is that AI-generated content — coordinated bot networks, deepfake "leaked" executive statements, AI-generated "panic" posts— could amplify a 2023-style social run by an order of magnitude.

Third, AI-driven flash events. The May 6, 2010, flash crash saw the Dow Jones Industrial Average fall 998.5 points (approximately 9%) intraday in roughly 36 minutes, triggered by a $4.1 billion algorithmic E-Mini sell program (U.S. CFTC & SEC, 2010). August 5, 2024, yen carry-trade unwind saw the TOPIX index lose 12% in a single day and the VIX briefly exceed levels not seen since COVID-19, with the S&P 500 falling 3.0% the same day (Lombardiet al., 2024). These episodes occurred without frontier-AI involvement. Future flash events involving correlated AI trading strategies —particularly if responding to AI-generated signals or news— could combine the speed of the 2010 crash with the scale of the 2024 carry unwind.

3.3. Credit and market risk

Credit-scoring AI is explicitly classified as a high-risk AI system under Annex III(5)(b) of the EU AI Act (Regulation (EU) 2024/1689), with the narrow exception of systems used for financial-fraud detection. This triggers the full suite of obligations in Articles 9–15 and 16. In the U.S., the CFPB's Circular 2023-03 (Consumer Financial Protection Bureau, 2023) requires that adverse-action notices specify the actual reasons for credit denial even when those denials are produced by complex algorithms: as then-Director Chopra stated, "creditors may not rely on the checklist of reasons provided in the sample forms… if those reasons do not specifically and accurately indicate the principal reason(s) for the adverse action." The CFPB's chatbot spotlight (2023) estimated that approximately 37% of the U.S. population interacted with a banking chatbot in 2022, projected to grow to over 40% by 2026.

Algorithmic trading is regulated in the EU via MiFID II RTS 6 (Commission Delegated Regulation (EU) 2017/589) and reinforced by an ESMA Supervisory Briefing of February 26, 2026, explicitly noting that algorithmic trading systems meeting the AI Act definition of "AI system" must comply with both regimes. In the U.S., the SEC's 2023 proposed rule on predictive data analytics (SEC, 2023) was withdrawn in June 2025; the SEC's first AI-washing enforcement actions against Delphia and Global Predictions (SEC, 2024) —yielding penalties of $225,000 and $175,000 respectively— represent a more limited but precedent-setting alternative.

Deepfake-enabled fraud has scaled rapidly. The Arup case (January 2024) involved 15 fraudulent transfers totaling HK$200 million (approx. US$25.6 million) in a single day, induced by a deepfake video conference featuring the AI-generated likenesses of the CFO and other senior staff (CNN, 2024; Hong Kong Police Force, 2024). Sumsub (2025) reported a 180% increase in sophisticated fraud in 2025 versus 2024, and Pindrop (2025) documented a 1,300% surge in deepfake fraud attempts in 2024, growing from one per month to seven per day on average. The FBI IC3 (2025) recorded $16.6 billion in total reported cyber-enabled losses in 2024, a 33% year-over-year increase, with $2.77 billion attributable to Business Email Compromise.

3.4. Governance, model risk, and explainability

The "black-box" problem is exacerbated, not solved, by the latest generation of models. ESMA's 2024 public statement on AI in retail investment services (ESMA, 2024) cited empirical research finding ChatGPT hallucinates in approximately 19.5% of responses — a rate that is operationally unacceptable for customer-facing financial-advice applications. The Bank of England (2024) found that "only a third of respondents describe themselves as having a complete understanding of the AI technologies they had implemented" and that 55% of all reported AI use cases have some form of automated decision-making, although only 2% are fully autonomous as of April 2025 (Bank of England, 2025).

Governance frameworks differ materially across jurisdictions. The NIST AI Risk Management Framework (NIST, 2023) is voluntary, technology-neutral, and cross-sectoral, organized around four core functions: GOVERN, MAP, MEASURE, MANAGE. NIST AI 600-1, the Generative AI Profile (Autio et al., 2024), adds 12 GenAI-specific risk categories including confabulation, harmful bias, and value-chain integration — but does not constitute a financial-services-specific profile. The EU AI Act's Chapter III Section 2 mandates a specific risk-management system (Art. 9), data governance (Art. 10), and explicit human oversight (Art. 14) for high-risk systems. EIOPA's Opinion of August 6, 2025 (EIOPA, 2025) applies similar governance principles to insurance.

3.5. Money laundering, fraud, and conduct risk

Generative-AI-enabled fraud is no longer hypothetical. The EBA and ECB jointly reported €4.2 billion in payment fraud across the European Economic Area in 2024, with manipulation of payers (social engineering) accounting for over half of fraudulent credit transfers (EBA & ECB, 2025). Spain's INCIBE managed 122,223 cybersecurity incidents in 2025, up 26% year-over-year, of which 34% of incidents affecting essential and important operators (NIS2 scope) hit the banking sector (INCIBE, 2026). Globally, FBI IC3 reported $16.6 billion in cyber-fraud losses in 2024 (FBI IC3, 2025), and Sophos (2024) found that 65% of financial-services organizations were hit by ransomware in 2024, with a median ransom payment of $2 million for those who paid.

The FS-ISAC AI Risk Working Group issued six white papers in February 2024 covering adversarial AI, defensive AI integration, generative AI vendor evaluation, and acceptable-use policies (FS-ISAC, 2024). These industry frameworks complement but do not substitute for regulatory action.

3.6. Concentration and oligopoly risk

The structural concentration of frontier AI is now well-documented. Synergy Research Group (2025) reports the Big Three hyperscalers (AWS at approx. 30%, Azure at approx. 21%, Google Cloud at approx. 13%) hold 63% of global enterprise cloud spending, with full-year 2024 cloud spend reaching approximately $330 billion. IoT

Analytics (2025) estimates NVIDIA held approximately 92% of the data-center GPU market in 2024, in a market worth $125 billion. In frontier models, the universe of "ASL-4-relevant" providers is even smaller: Anthropic, OpenAI, Google DeepMind, with secondary players including xAI, Meta AI, and Mistral.

For banking specifically, Breeden (2024) reported that 44% of third-party AI models used by UK financial firms come from the top three providers, up from 18% in 2022. The Treasury's December 2024 report on AI in financial services (U.S. Department of the Treasury, 2024c) flagged "third-party concentration risk (few firms dominating advanced model market)" as a principal concern. DORA's Critical ICT Third-Party Provider (CTPP) regime (Arts. 31–44; Regulation (EU) 2022/2554), supplemented by Commission Delegated Regulation (EU) 2024/1502 of 22 February 2024, allows the European Supervisory Authorities to designate critical third-party providers and impose fines of up to 1% of average daily worldwide turnover for non-compliance with recommendations. As of mid-2026, the CTPP designations have not been finalized.

4. POSITIVE EFFECTS OF AI IN BANKING: THE COUNTERWEIGHT

A balanced assessment must register that AI also makes the financial system more resilient on multiple dimensions. The BIS (2024) reported that approximately 70% of financial-services firms are using AI to enhance cash-flow predictions, improve liquidity management, fine-tune credit scoring, and improve fraud detection. The U.S. Treasury (Oct. 17, 2024, press release, cited in Bowman, 2024) credited machine-learning AI fraud-detection tools with preventing and recovering over $4 billion in fraud in fiscal year 2024, including approximately $1 billion in Treasury check fraud.

In credit underwriting, AI enables the use of alternative data to extend credit to thin-file consumers — though the CFPB and EU AI Act both insist on adverse-action transparency. In AML, BIS Innovation Hub's Project Aurora and similar initiatives demonstrate that ML-based anomaly detection can outperform rules-based systems. In customer service, Bank of America's Erica reached 2.5 billion cumulative interactions by end-2024, with 676 million in 2024 alone serving 20 million unique users (Bank of America, 2025). BBVA reported that its OpenAI deployment saved approximately three hours per week per employee on routine tasks, with over 80% daily engagement, leading to expansion from 3,300 to 11,000 licenses in 2025, and to all 120,000 employees across 25 countries in December 2025 (OpenAI, 2025; BBVA, 2025).

In risk modeling, EBA's follow-up report on machine learning for IRB models (EBA, 2023) is cautiously permissive, allowing ML in PD model development subject to explainability and stability constraints. The ECB's revised Internal Models Guide (ECB, 2025c) similarly opens a regulated path. The Banco de España is recruiting 25 AI and data-science professionals to develop supervisory AI tools (Banco de España, 2025b). The ECB has deployed supervisory AI tools including Athena (document search), Virtual Lab, Delphi (early risk detection), Medusa, and Heimdall (fit-and-proper assessments) (ECB Banking Supervision, 2025).

The pivotal point is that the same capability advance that creates dual-use offensive risk (Mythos finding 271 Firefox vulnerabilities) also creates defensive benefit (the same vulnerabilities being patched). The question is whether the net effect on financial stability is positive — and the answer depends crucially on the policy and supervisory framework that governs disclosure, access, and deployment.

5. COMPARATIVE REGULATORY LANDSCAPE: SPAIN, THE EU, AND THE UNITED STATES

5.1. The United States: Principles-based, increasingly fragmented

The U.S. framework is layered and partly contradictory in 2026. Executive Order 14110 (October 30, 2023) directed over 50 federal entities to undertake more than 100 actions, including the Treasury's March 2024 AI cybersecurity report (U.S. Department of the Treasury, 2024a) and NIST's AI 600-1 GenAI Profile (Autio et al., 2024). EO 14110 was rescinded by EO 14148 on January 20, 2025, and EO 14179 of January 23, 2025, declared

it U.S. policy "to sustain and enhance America's global AI dominance" (Executive Order 14179, 2025). A further Executive Order of December 11, 2025 ("Eliminating State Law Obstruction of National Artificial Intelligence Policy") aims to preempt state AI laws.

For banking, the principal microprudential instrument is SR 26-2 (Federal Reserve Board et al., 2026), which supersedes SR 11-7 (Board of Governors of the Federal Reserve System & OCC, 2011) and SR 21-8. As discussed in Section 3.1, generative and agentic AI are explicitly out of scope. Third-party risk is governed by SR 23-4 (the 2023 Interagency Guidance on Third-Party Relationships). The CFPB's chatbot spotlight and Circular 2023-03 govern consumer-facing AI. The SEC has used existing antifraud authorities to pursue "AI washing" (SEC, 2024). NIST's Center for AI Standards and Innovation (CAISI), renamed from US AISI in June 2025, has signed voluntary pre-deployment evaluation agreements with Anthropic, OpenAI, Google, Microsoft, and xAI (NIST, 2024).

The most consequential supervisory action in the period studied was the April 7–8, 2026 Treasury–Fed–bank-CEO meeting on Mythos, which Vice Chair Bowman confirmed in her May 1, 2026, speech (Bowman, 2026). Bowman also disclosed that she chairs the FSB Standing Committee on Supervisory and Regulatory Cooperation and that the consultation report on sound AI practices would be released for comment in Q3 2026.

5.2. The European Union: The most prescriptive horizontal framework

The EU framework is built on three pillars. First, the EU AI Act (Regulation (EU) 2024/1689) entered into force on August 1, 2024, with staggered application: prohibitions and AI literacy obligations from February 2, 2025; GPAI rules, governance, and penalties from August 2, 2025; main high-risk obligations from August 2, 2026; and Article 6(1) safety-component rules from August 2, 2027. Credit-scoring AI is explicitly high-risk under Annex III(5)(b). General-Purpose AI models trained with cumulative compute exceeding 10^{25} FLOPs are classified as having systemic risk (Art. 51). Penalties run up to €35 million or 7% of worldwide annual turnover for Article 5 prohibitions, €15 million or 3% for other obligations.

Second, DORA (Regulation (EU) 2022/2554) applied from January 17, 2025, and provides the operational-resilience backbone. Its five pillars — ICT risk management, incident management and reporting, digital operational resilience testing including TLPT every three years (Arts. 26–27), ICT third-party risk including CTPP designation (Arts. 31–44), and information-sharing — apply to approx. 20 categories of financial entities plus ICT third-party providers. The TLPT RTS was published as Commission Delegated Regulation (EU) 2025/1190 on June 18, 2025. The first DORA-TLPT cycle for designated entities must be completed by January 17, 2028.

Third, the SSM supervisory framework. The ECB's SSM Supervisory Priorities 2026–2028 (ECB, 2025d) note that operational and ICT risk continue to receive the worst average SREP scores. The ECB's 2024 cyber resilience stress test covered 109 directly supervised banks, with 28 subjected to deep-dive testing including actual IT recovery (ECB, 2024b). The ECB Macroprudential Bulletin of February 2025 (ECB, 2025a) extended the analysis to macroprudential perspectives, identifying three contagion channels for cyber incidents: operational, financial, and confidence.

The ESRB Advisory Scientific Committee Report No. 16 (ESRB, 2025), published December 4, 2025, provides the most rigorous European systemic-risk analysis. Frank Elderson's May 13, 2026 newsletter on Mythos (NL Times, 2026) and Christine Lagarde's repeated emphasis on AI risk (e.g., Lagarde, 2025) demonstrate the high political salience of AI within the ECB.

5.3. Spain: The first national AI supervisory agency in the EU

Spain has pursued a strategy of front-running EU obligations. AESIA (Agencia Española de Supervisión de Inteligencia Artificial) was created by Real Decreto 729/2023 of 22 August 2023 (Boletín Oficial del Estado, 2023), making Spain the first EU member state to establish a national AI supervisory agency — eleven months before

the EU AI Act entered into force. Inspection authority for prohibited AI practices began February 2, 2025, with full sanctioning powers from August 2, 2025. Staff numbers reached approximately 30 by late 2025.

The Banco de España maintains banking supervisory responsibility (within the SSM) and publishes its Informe de Estabilidad Financiera biannually. The autumn 2025 IEF (Banco de España, 2025a) reports a Spanish banking sector with ROE of 14.6% and CET1 of 13.8% as of June 2025, with cyber and hybrid risks treated within the geopolitical-risk section. The bank has launched a transversal AI working group, while noting that the Banco de España has only 5% of the technological resources of larger eurosystem peers despite holding 12% of ECB capital.

The CNMV addressed AI in capital markets in its Plan de Actividades 2024 (CNMV, 2024) and has hired 76 new staff for MiCA and DORA supervision. Its AI activities include exploring generative AI in investment services, sandbox analysis, and internal-process improvement.

Spain's broader policy framework includes the 2024 update to ENIA (Estrategia de Inteligencia Artificial) with €1.5 billion in 2024–2025 funding (in addition to €600 million from 2021–2023), and a portfolio of initiatives including ALIA (Spanish-language open generative foundation model), Quantum Spain, and IA en Cadenas de Valor (Gobierno de España, 2024).

TABLE 2. REGULATORY COMPARISON MATRIX — SPAIN, EU, U.S.

Dimension	Spain	EU (SSM, AI Act, DORA)	U.S. (Fed, OCC, FDIC, CFPB, SEC, NIST)
Horizontal AI law	EU AI Act (direct effect); AESIA as supervisor	EU AI Act 2024/1689 (in force Aug 1, 2024)	None; EO 14179 prefers deregulation
AI banking-specific	SR 26-2 not applicable; AI Act applies	EU AI Act Annex III(5)(b) credit scoring high-risk; ECB internal-models guide (2025)	SR 26-2 (Apr 2026) — excludes gen AI/agentic AI
Operational resilience	DORA via direct effect	DORA (Reg. 2022/2554) — applies from Jan 17, 2025	SR 23-4 (2023 third-party guidance); voluntary FFIEC IT Handbook
Cyber stress testing	TIBER-EU/TIBER-ES; CBE participation	TLPT every 3 years (Arts. 26–27) by 2028; ECB 109-bank cyber stress test 2024	CBEST equivalent at state level (NYDFS); no horizontal federal cyber stress test
GPAI / frontier model rules	EU AI Act direct effect	Art. 51 systemic-risk GPAI (10^{25} FLOPs); GPAI Code of Practice (July 2025)	Voluntary CAISI MOUs with Anthropic, OpenAI, Google, Microsoft, xAI
Penalties	Per EU AI Act	Up to €35M / 7% worldwide turnover	SEC, OCC, CFPB, FRB ad hoc; no AI-specific scheme
National AI agency	AESIA (since Aug 2023; operational Feb 2025)	European AI Office (since Jan 2024)	NIST CAISI (renamed from U.S. AISI June 2025)

6. CHANNELS OF CONTAGION AND SYSTEMIC AMPLIFICATION

A model of how AI-related shocks propagate through the financial system must integrate the operational, financial, and confidence channels identified by the ESRB (2020) and ECB (2025a) with the AI-specific amplifying features.

Stage 1: The shock origin. A frontier model —wielded by an adversary, a state actor, or even an inadequately-supervised legitimate user— identifies an exploit in a critical software dependency of one or more financial institutions. Calibration: Mythos found 271 Firefox vulnerabilities in a single evaluation pass (The Next Web, 2026) and reports thousands of high-severity vulnerabilities across major OSes and browsers (Anthropic, 2026a).

Stage 2: The operational layer. The exploit is weaponized; banks experience disruption of core systems, payment rails, or customer-facing channels. Calibration: ICBC's November 2023 ransomware attack disrupted over $9 billion in Treasury settlements (Treasury, 2024b); the July 19, 2024, CrowdStrike outage affected approx. 8.5 million Windows devices (Microsoft, 2024), with Delta alone losing $500 million (Delta Air Lines SEC 8-K, 2024).

Stage 3: The financial layer. Banks unable to process payments or honor obligations begin to default on inter-bank exposures or trigger margin calls. Counterparties impose tighter terms; liquidity hoarding begins. Calibration: SVB's negative cash balance at end of March 9 was approximately $958 million (CA DFPI, 2023); the August 5, 2024, yen carry-trade unwind saw forward carry positions of approx. $160 billion unwind globally (Lombardi et al., 2024).

Stage 4: The confidence layer. Social media and AI-amplified communications spread information about the disruption. Depositors and counterparties respond. Calibration: 6,528 SVB "run" tweets posted between March 8–13, 2023; banks in the top tercile of Twitter exposure lost 6.6 percentage points more stock value (Cookson et al., 2025).

Stage 5: Macroprudential transmission. If banks experience correlated runs and have correlated AI-driven asset positions (e.g., similar model-driven hedges), simultaneous unwinding produces market dislocations. Calibration: the May 6, 2010, flash crash saw $4.1 billion in algorithmic sales trigger a 9% DJIA drop in 36 minutes (U.S. CFTC & SEC, 2010).

AI amplifies each stage: the speed feature compresses Stages 1–2; the uniformity feature correlates Stage 3 across institutions; the monitoring challenges and speed features amplify Stage 4; the concentration feature converts a per-firm shock into a sector-wide one.

7. SCENARIO ANALYSES: FOUR CALIBRATED SIMULATIONS

7.1. Scenario A: AI-driven cyberattack on a G-SIB

Calibration baseline: IBM Cost of a Data Breach 2024 reports a financial-services average breach cost of $6.08 million, with mega-breach (>50 million records) average cost of $375 million (IBM Security, 2024). Mean detection plus containment time: 258 days. Sophos (2024) reports a 65% ransomware-attack rate in financial services with median ransom of $2 million and mean recovery cost of $2.58 million. The IMF (2024 GFSR April) projected once-in-a-decade cyber losses for a financial firm of $2.2 billion.

Adverse scenario: A Mythos-class autonomous-attack actor (whether state-sponsored or criminal with frontier-model access via leak or model extraction) deploys multi-target, multi-vector attacks against a G-SIB's core banking systems and three of its major vendors. Calibrated parameters: 8.5 million affected endpoints (CrowdStrike analog); 8–24 hour service disruption (CrowdStrike was reverted in 78 minutes, but cleanup took days); concurrent attack on Treasury settlement infrastructure (ICBC analog) disrupting >$9 billion in transactions; ransom demand at the 95th percentile of Sophos 2024 financial-services data.

Projected impact (illustrative, grounded in calibration points): Direct losses including service disruption, customer remediation, and recovery costs scale toward IMF's $2.2 billion once-in-a-decade estimate. Indirect losses —regulatory fines (Capital One $80M OCC fine; OCC, 2020), class-action settlements ($190M; Capital

One Settlement Administration, 2022), and lost business— could match or exceed direct losses. Recovery time materially exceeds the ECB cyber stress test's tolerable threshold for many institutions (ECB, 2024b).

7.2. Scenario B: AI-amplified deposit run

Calibration baseline: SVB lost $42 billion (approx. 25% of $166 billion total deposits) on March 9, 2023, with another $100 billion in withdrawal requests pending for March 10 (Federal Reserve Board, 2023). Year-end 2022 uninsured deposits were $151.6 billion or 93.8% of total (CA DFPI, 2023). Cookson et al. (2025) demonstrate a 4.3-percentage-point excess stock decline per standard deviation of pre-existing Twitter exposure, rising to 6.6 percentage points for top-tercile-exposed banks.

Adverse scenario: A regional bank (approx. $200 billion assets, approx. 60% uninsured deposits) experiences a coordinated AI-generated misinformation campaign — including deepfake video of "leaked" internal executive panic, AI-generated mass social media posts in multiple languages, and synthetic news articles. The campaign is launched outside business hours and amplified by coordinated bot networks across X, TikTok, WhatsApp, and Telegram.

Projected impact. Calibrated to SVB but compressed: SVB experienced 25% deposit outflows in one trading session; the projection assumes 25–40% outflows in 12 hours given (i) the maturation of digital-banking infrastructure since 2023, (ii) the amplification factor of AI-generated content, and (iii) the precedent set by SVB. The Bank of England (2025) noted that 70% of UK Systemic Risk Survey respondents in 2024 H1 cited cyberattacks as a UK financial system risk, suggesting widespread awareness of this vulnerability among supervisors.

7.3. Scenario C: Algorithmic herding flash event

Calibration baseline: May 6, 2010: DJIA fell 998.5 points (approx. 9%) intraday in 36 minutes; $4.1 billion algorithmic E-Mini sell program was the trigger; total May 6 volume was 19.4 billion shares (U.S. CFTC & SEC, 2010). August 5, 2024: TOPIX -12% in one day (BIS Bulletin 90; Lombardi et al., 2024), Nikkei -12.4%, VIX briefly above 60; approx. $160 billion in hedge fund forward carry positions; recovery within one week. IMF (2024) reports algorithmic trading at approx. 70% of U.S. equities and approx. 50% of U.S. futures.

Adverse scenario: A correlated AI-driven trading strategy —for example, multiple banks and hedge funds using similar large-language-model-based news-sentiment overlays on positions in a concentrated sector— receives an AI-generated false news shock (perhaps generated by an adversary). Models simultaneously trigger selling.

Projected impact: The scenario combines the speed of 2010 (algorithms reacting in milliseconds) with the scale of 2024 (cross-border carry-trade-like positions). If markets respond similarly to the 2024 episode, single-day equity declines of 10–12% in affected indices are plausible, with VIX exceeding 60. Circuit breakers, redesigned post-2010, limit the worst outcomes — but the IMF (2024) explicitly recommended recalibration of circuit breakers "in light of potentially rapid AI-driven price moves."

7.4. Scenario D: Third-party AI provider outage or cascade

Calibration baseline: CrowdStrike (July 19, 2024) affected approx. 8.5 million Windows devices, with approx. 24,000 CrowdStrike customers including approx. 60% of Fortune 500; Delta alone reported approx. $500 million loss; Parametrix estimated $5.4 billion total losses to top 500 U.S. companies excluding Microsoft (Parametrix, 2024). Cloud-service disruptions globally totaled 205.3 hours of outage in 2023, up from 133.5 hours in 2022 (Parametrix, cited in IMF, 2024).

Adverse scenario: A major AI provider (Anthropic, OpenAI) or hyperscaler (AWS, Azure, GCP) experiences a 12–48-hour disruption — through misconfigured update (CrowdStrike analog), targeted attack (Mythos-class

adversary), or model failure (data poisoning at training). Given that 44% of UK FS third-party AI models come from the top three providers (Breeden, 2024) and that AWS/Azure/GCP collectively host approx. 63% of enterprise cloud workloads (Synergy Research Group, 2025), the disruption cascades across multiple G-SIBs simultaneously.

Projected impact: Direct service-disruption costs scale with CrowdStrike-equivalent estimates. Critically, correlated disruption across multiple G-SIBs converts what would be a per-firm operational incident into a sector-wide one. The DORA CTPP regime is the principal mitigation, but Critical Third-Party Provider designations are not yet finalized as of mid-2026, and even after designation, the regime relies on recommendations and oversight rather than direct prudential authority.

TABLE 3. SCENARIO SUMMARY

Scenario	Calibration source	Headline impact (illustrative)	Key mitigation
A. AI-driven G-SIB cyberattack	IBM 2024; Sophos 2024; ICBC 2023; IMF 2024	$1–2B+ direct loss; multi-day disruption	DORA TLPT; SR 26-2; ECB cyber stress framework
B. AI-amplified deposit run	SVB Mar 2023; Cookson et al., 2025	25–40% deposit outflows in 12 hours	Real-time liquidity reporting; FDIC reform
C. Algorithmic herding flash event	May 6, 2010 crash; Aug 5, 2024 unwind	-10% to -12% equity decline; VIX >60	Recalibrated circuit breakers; margin review
D. AI provider/hyperscaler cascade	CrowdStrike Jul 2024; cloud outage 205.3 hrs (2023)	Multi-G-SIB simultaneous disruption	DORA CTPP regime; cloud-exit drills

8. SPAIN-SPECIFIC DEEP DIVE

8.1. Cyber posture and incident exposure

INCIBE managed 122,223 cybersecurity incidents in 2025, a 26% year-over-year increase (INCIBE, 2026). Of the 401 incidents affecting essential and important operators under NIS2 scope, banking accounted for 34% — the largest sectoral share. The 2024 balance recorded 97,348 incidents and 341 NIS2-scope incidents.

The Spanish cybersecurity industry generated €6,351 million in revenue in 2024 (+70% since 2020), with 164,761 cybersecurity professionals — making Spain the EU's fourth-largest cybersecurity market (12% of continental revenue). However, Spain missed the October 17, 2024, NIS2 transposition deadline, with the European Commission issuing a reasoned opinion in May 2025 and the implementing law still not published as of mid-2026.

8.2. Cross-border dependencies

Spain's banks are uniquely exposed to cross-border AI/cyber transmission. Santander has G-SIB designation (Banco de España, 2025a). Its 2024 revenue mix shows Brazil at 21.6%, Spain 17.4%, U.S. 12.3%, U.K. 11.2%, Mexico 10.2%, and Poland 6.3%. Santander Bank N.A. has $102 billion in U.S. assets and 1.8 million U.S. customers.

8.3. Supervisory capacity

The Banco de España has approximately 3,475 professionals, compared to 6,968 at the Banca d'Italia, 8,958 at the Banque de France, and 10,255 at the Bundesbank. Governor Escrivá has publicly noted that the Banco de España has 5% of the technological resources of larger Eurosystem peers despite holding 12% of ECB capital. AESIA, with approx. 30 staff in 2025, faces a similar scale challenge relative to its EU peers and the breadth of its statutory mandate. The Financial Stability Board's Peer Review of Spain (November 18, 2025) recommended developing a comprehensive sectoral cyber-threat landscape, national third-party risk analysis, and streamlined incident notification.

9. POLICY RECOMMENDATIONS

The Mythos affair has revealed that the existing regulatory architecture, while substantially developed, has critical gaps. Five recommendations follow.

9.1. For microprudential supervisors

First, revisit the exclusion of generative and agentic AI from model risk management. SR 26-2's footnote 3, which removes gen AI and agentic AI from scope (Federal Reserve Board et al., 2026), reflects a recognition that traditional MRM principles do not map well to LLMs. But the alternative cannot be regulatory silence. Supervisors should issue interim guidance —building on NIST AI 600-1 (Autio et al., 2024), the EBA Follow-up on ML in IRB (EBA, 2023), and EIOPA's Opinion (EIOPA, 2025)— that establishes minimum expectations for gen AI use in regulated activities. The EU's path under Articles 9–15 of the AI Act provides a usable template.

Second, harmonize cyber stress testing across jurisdictions. The ECB's 2024 cyber stress test (109 banks, 28 in deep dive; ECB, 2024b) sets the operative European standard; the Bank of England's CBEST and SIMEX programmes (Bank of England, 2022, 2023) provide the U.K. analog. The U.S. should develop a horizontal federal equivalent —building on NYDFS state-level work— to be exercised annually.

9.2. For macroprudential authorities

Third, designate critical AI providers under DORA-like regimes globally. DORA's CTPP regime (Arts. 31–44; Commission Delegated Regulation (EU) 2024/1502) is the most developed framework. CTPP designations should be finalized urgently and explicitly extended to cover frontier AI labs alongside cloud hyperscalers. The U.S. should establish a parallel designation regime, potentially through Financial Stability Oversight Council action. The FSB's 2026 sound practices report should provide a converged international template.

Fourth, incorporate AI-amplified run scenarios into liquidity and resolution frameworks. Cookson et al.'s (2025) finding of 4.3–6.6 percentage points of excess loss from social-media exposure during the SVB run is now empirically validated. Liquidity ratios calibrated to pre-2023 outflow assumptions are obsolete. The FDIC, Federal Reserve, ECB SSM, and Banco de España should jointly update outflow assumptions in LCR and resolution planning to reflect digital-age run dynamics.

9.3. For industry

Fifth, standardize AI vendor risk management and adopt the FS-ISAC framework as a floor. FS-ISAC's six AI white papers (FS-ISAC, 2024) —particularly Generative AI Vendor Evaluation and Qualitative Risk Assessment— provide a practical, industry-developed standard. Banks should adopt these (or equivalent) as floor standards, integrated with NIST AI RMF (NIST, 2023) and the EU AI Act's conformity assessment requirements.

9.4. For frontier AI labs

Sixth, formalize responsible disclosure protocols for AI-discovered financial-infrastructure vulnerabilities. Anthropic's bespoke arrangement with the FSB —agreeing to brief, restricting wider distribution at White House request (Financial Times, 2026), maintaining the approx. 40-organization Glasswing partner list— improvised a path that should be institutionalized. A protocol should specify: (i) which national institutes (US CAISI, UK AISI) have standing access to pre-deployment evaluations; (ii) which sectoral bodies (FS-ISAC, ECB SSM, Banco de España) receive timely briefings on capability advances; (iii) how access to defensive capabilities is broadened symmetrically (the Mythos asymmetry —40 organizations with access, thousands without— is unstable); and (iv) how disclosure timelines balance patch-time and exploit-time.

Seventh, expand independent pre-deployment evaluation. METR's task-length benchmarks (Kwa et al., 2025), Apollo Research's scheming evaluations (Apollo Research, 2024), UK AISI's cyber-capability evaluations (UK AI Security Institute, 2026a), and XBOW's offensive-security benchmarks (XBOW, 2026) collectively provide a maturing evaluation ecosystem. Anthropic's RSP (Anthropic, 2025), OpenAI's Preparedness Framework v2 (OpenAI, 2025), and Google DeepMind's Frontier Safety Framework v3 (Google DeepMind, 2025) provide industry-side analogs. Pre-deployment evaluation should become binding rather than voluntary for frontier models offered to systemic financial institutions.

10. CONCLUSION

The May 2026 Mythos affair is, on the surface, a story about one frontier AI laboratory engaging with one global standard-setting body about cyber vulnerabilities in financial infrastructure. Beneath the surface, it is a story about three structural shifts that the financial-stability community must now confront.

The first shift is in the nature of cyber risk itself. When a frontier model can find thousands of high-severity vulnerabilities in days and write working exploits with an 83.1% pass rate on the first attempt (Anthropic, 2026a), the time-to-exploit collapses below the time-to-patch — and the equilibrium assumptions underlying decades of operational-resilience policy become outdated. The IMF's projection that once-in-a-decade cyber losses for a financial firm could reach $2.2 billion (IMF, 2026) may itself be conservative by 2027 standards.

The second shift is in the regulatory perimeter. A handful of frontier AI labs and cloud hyperscalers —Anthropic, OpenAI, Google DeepMind, AWS, Azure, Google Cloud, NVIDIA— are now operationally as central to financial stability as the largest banks, but they sit outside the prudential perimeter. The DORA CTPP regime provides the most developed framework for closing this gap, but it remains in early implementation. The U.S. has no equivalent. The FSB's planned sound practices report is the most consequential international policy initiative in this space, and its design will shape the rules of the next decade.

The third shift is in the politics of AI capability. Anthropic's agreement not to distribute Mythos more widely at the request of the White House (Financial Times, 2026) sets a precedent that frontier-AI capability is now a matter of national-security policy as well as commercial policy. As Mistral CEO Arthur Mensch noted in May 2026, "it is impossible to have the source code of the French army checked by Mythos" — and the implication of "irreversible dependency" extends to financial infrastructure as well (NL Times, 2026). The European Commission, the Bundesbank, and the Banco de España have all called for broader access. A regime in which Spanish banks must rely on a U.S. lab's discretionary disclosure to a U.K.-chaired international body is unstable.

For Spain specifically, the policy implications are sharp. The Banco de España, CNMV, and AESIA together possess most of the supervisory authorities they need, but with relatively limited resources and an unfinished NIS2 transposition. Santander's G-SIB status and BBVA's deep OpenAI integration —extending to all 120,000 employees in 25 countries by end of 2025 (OpenAI, 2025)— make Spanish banks unusually exposed to both upside and downside scenarios. The Spanish banking system's high concentration (CR3 approx. 72%; World

Bank) means that a Mythos-class incident affecting one or two large institutions could plausibly become a systemic event.

The deepest lesson of the Mythos moment is that financial stability has acquired a new dependency on the responsible behavior of a small number of non-financial actors. The traditional financial-stability framework assumed that the relevant systemically important entities were banks (with derivatives clearinghouses, CCPs, and a few others added over time). The Mythos affair demonstrates that frontier AI labs are now in this set — not because they take deposits, but because their decisions about model capability, model release, and disclosure protocols can materially affect the operational continuity of every bank in the world.

The Financial Stability Board, the Federal Reserve, the European Central Bank, and the Banco de España will need new tools —or, more accurately, fresh applications of existing tools— to manage this new dependency. The first generation of those tools is being built right now, in the briefing rooms of the FSB, in the consultation drafts of DORA implementing acts, and in the model cards of Claude Opus 5 and GPT-6. Whether they are built well, fast, and in coordination across jurisdictions is the central question of financial-stability policy for the remainder of the decade.

References

Anthropic. (2025). Responsible Scaling Policy. https://www.anthropic.com/rsp

Anthropic. (2026a). Project Glasswing: Introducing Claude Mythos Preview. https://www.anthropic.com/glasswing

Anthropic. (2026b, February). Anthropic raises Series G at $380 billion post-money valuation [Press release].

Apollo Research. (2024). Scheming evaluations of frontier models. https://www.apolloresearch.ai

Autio, E., Stanley, K., & National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

BAE Systems. (2016). Two bytes to $951m: An investigation into the SWIFT/Bangladesh Bank cyber-heist. BAE Systems Threat Research Blog.

Banco de España. (2025a). *Informe de Estabilidad Financiera — Otoño 2025*. https://www.bde.es/f/webbe/Secciones/Publicaciones/InformesBoletinesRevistas/InformesEstabilidadFinancera/25/IEF_Otono2025.pdf

Banco de España. (2025b, April). Plan Estratégico 2024–2027. Banco de España.

Banco Santander. (2025). Annual Report 2024. https://www.santander.com

Bank of America. (2025, February). Digital interactions by BofA clients surge to over 26 billion, up 12% year-over-year [Press release]. https://newsroom.bankofamerica.com

Bank of England. (2022). CBEST: Threat intelligence-led testing. https://www.bankofengland.co.uk

Bank of England. (2023). SIMEX 2022: Sector-wide cyber simulation exercise. https://www.bankofengland.co.uk

Bank of England. (2024). *Machine learning in UK financial services 2024*. Bank of England and Financial Conduct Authority.

Bank of England. (2025, April). *Financial Stability in Focus: Artificial intelligence in the financial system*. https://www.bankofengland.co.uk/financial-stability-in-focus/2025/april-2025

Bank of England. (2026, April). Speech by Andrew Bailey at Columbia University [Speech]. https://www.bankofengland.co.uk

Bloomberg. (2026a, April 22). Bank of England's CMORG places Mythos on operational resilience agenda. Bloomberg News.

Bloomberg. (2026b, May). OpenAI valuation update. Bloomberg News.

Board of Governors of the Federal Reserve System. (2023). Review of the Federal Reserve's supervision and regulation of Silicon Valley Bank. https://www.federalreserve.gov/publications/files/svb-review-20230428.pdf

Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR 11-7 / OCC Bulletin 2011-12). https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Bowman, M. W. (2024, October 17). Remarks on artificial intelligence in financial services [Speech]. Board of Governors of the Federal Reserve System.

Bowman, M. W. (2026, May 1). Artificial intelligence in the financial system [Speech]. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/newsevents/speech/bowman20260501a.htm

Breeden, S. (2024, November 4). Engaging with the machine: AI and financial stability [Speech].

Bank for International Settlements. https://www.bis.org/review/r241104j.htm

California Department of Financial Protection and Innovation. (2023). Review of DFPI's oversight and regulation of Silicon Valley Bank. https://dfpi.ca.gov

Capital One Settlement Administration. (2022). Notice of class action settlement: In re Capital One Consumer Data Security Breach Litigation.

Cisa & Federal Bureau of Investigation. (2023, June 7). #StopRansomware: CL0P ransomware gang exploits CVE-2023-34362 MOVEit vulnerability (Alert AA23-158A). U.S. Cybersecurity and Infrastructure Security Agency.

CNBC. (2025, October). Microsoft increases OpenAI investment. CNBC.

CNBC. (2026, April 10). Powell, Bessent discussed Anthropic's Mythos AI cyber threat with major U.S. banks. https://www.cnbc.com

CNMV (Comisión Nacional del Mercado de Valores). (2024). Plan de Actividades 2024. https://www.cnmv.es

CNMV (Comisión Nacional del Mercado de Valores). (2025, October 16). Resultado de la OPA de BBVA sobre Banco Sabadell [Press release]. https://www.cnmv.es

Coalition Greenwich. (2024). U.S. equity trading: Algorithmic and program trading trends. Greenwich Associates.

Consumer Financial Protection Bureau. (2023, September 19). Circular 2023-03: Adverse action notification requirements and the proper use of CFPB's sample forms when using artificial intelligence or complex credit models. https://www.consumerfinance.gov

Cookson, J. A., Fox, C., Gil-Bazo, J., Imbet, J. F., & Schiller, C. (2025). *Social media as a bank run catalyst.* Federal Reserve Bank of Cleveland Financial Stability Conference Papers. https://www.clevelandfed.org

Delta Air Lines. (2024, August). Form 8-K: Update on July 19, 2024 CrowdStrike-related outage. U.S. Securities and Exchange Commission.

EBA (European Banking Authority). (2023). Follow-up report on machine learning for IRB models. https://www.eba.europa.eu

EBA & ECB (European Central Bank). (2025). Joint report on payment fraud in 2024. https://www.eba.europa.eu

ECB (European Central Bank). (2024a, January 17). One step ahead: Protecting the cyber resilience of financial infrastructures [Speech]. https://www.ecb.europa.eu/press/key/date/2024/html/ecb.sp240117~3e839b396f.en.html

ECB (European Central Bank). (2024b, July 26). ECB concludes cyber resilience stress test [Press release]. https://www.bankingsupervision.europa.eu/press/pr/date/2024/html/ssm.pr240726~06d5776a02.en.html

ECB (European Central Bank). (2025a, February). Cyber resilience stress testing from a macroprudential perspective. ECB Macroprudential Bulletin. https://www.ecb.europa.eu

ECB (European Central Bank). (2025b). Speeches by Frank Elderson and Christine Lagarde on AI in the financial system. https://www.ecb.europa.eu

ECB (European Central Bank). (2025c). Guide to internal models (Revised). https://www.bankingsupervision.europa.eu

ECB (European Central Bank). (2025d, November). SSM supervisory priorities 2026–2028. https://www.bankingsupervision.europa.eu/framework/priorities/html/ssm.supervisory_priorities202511.en.html

ECB Banking Supervision. (2025, October 14). Artificial intelligence and supervision: Innovation with caution [Speech]. https://www.bankingsupervision.europa.eu

EIOPA (European Insurance and Occupational Pensions Authority). (2025, August 6). Opinion on artificial intelligence governance and risk management. https://www.eiopa.europa.eu

Emsisoft. (2023). MOVEit zero-day campaign: Tracking the impact. Emsisoft Threat Research.

ESMA (European Securities and Markets Authority). (2024). Public statement on the use of artificial intelligence in the provision of retail investment services. https://www.esma.europa.eu

ESMA (European Securities and Markets Authority). (2026, February 26). Supervisory briefing on algorithmic trading and AI systems. https://www.esma.europa.eu

ESRB (European Systemic Risk Board). (2020). Systemic cyber risk. Publications Office of the European Union.

ESRB Advisory Scientific Committee. (2025, December 4). Artificial intelligence and systemic risk (Report No. 16). European Systemic Risk Board. https://www.esrb.europa.eu

European Parliament & Council of the European Union. (2017). Commission Delegated Regulation (EU) 2017/589 of 19 July 2016 supplementing Directive 2014/65/EU with regard to regulatory technical standards specifying the organisational requirements of investment firms engaged in algorithmic trading [MiFID II RTS 6]. Official Journal of the European Union, L 87.

European Parliament & Council of the European Union. (2022). Regulation (EU) 2022/2554 on digital operational resilience for the financial sector [DORA]. Official Journal of the European Union, L 333.

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Evident Insights. (2025). Evident AI Index 2025. https://evidentinsights.com

Executive Order 14110. (2023, October 30). Safe, secure, and trustworthy development and use of artificial intelligence. 88 Fed. Reg. 75191 (rescinded January 20, 2025).

Executive Order 14148. (2025, January 20). Initial rescissions of harmful executive orders and actions. White House.

Executive Order 14179. (2025, January 23). Removing barriers to American leadership in artificial intelligence. White House.

FBI Internet Crime Complaint Center. (2025). 2024 Internet Crime Report. Federal Bureau of Investigation. https://www.ic3.gov

Federal Reserve Board, FDIC, & OCC. (2026, April 17). SR 26-2: Revised guidance on model risk management [Supervisory letter]. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm

Financial Times. (2026, May 18). Anthropic to brief financial regulators on cyber flaws exposed by Mythos.

FSB (Financial Stability Board). (2024, November). The financial stability implications of artificial intelligence. https://www.fsb.org/uploads/P14112024.pdf

FSB (Financial Stability Board). (2025, October). Monitoring adoption of artificial intelligence and related vulnerabilities in the financial sector. https://www.fsb.org

FSB (Financial Stability Board). (2025, November 18). Peer review of Spain. https://www.fsb.org/uploads/P181125.pdf

FSB (Financial Stability Board). (2026). Sound practices for the use of AI in the financial system [Consultation draft].

FS-ISAC. (2024). Building cryptographic agility in the financial sector and additional AI risk papers. Financial Services Information Sharing and Analysis Center.

Gobierno de España. (2024). Estrategia Nacional de Inteligencia Artificial 2024 (ENIA). Ministerio para la Transformación Digital y de la Función Pública.

Gobierno de España. (2023, August 22). Real Decreto 729/2023, de 22 de agosto, por el que se aprueba el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial. Boletín Oficial del Estado, 203.

Google DeepMind. (2025). Frontier Safety Framework, version 3. https://deepmind.google

Hong Kong Police Force. (2024, February). Statement on Arup deepfake fraud case.

IBM Security. (2024). Cost of a data breach report 2024. IBM Corporation. https://www.ibm.com/reports/data-breach

IMF (International Monetary Fund). (2024, October). *Global financial stability report: Chapter 3 — Advances in artificial intelligence: Implications for capital market activities.* International Monetary Fund. https://www.imf.org

IMF (International Monetary Fund). (2026). *Global financial stability report* (April). International Monetary Fund.

INCIBE (Instituto Nacional de Ciberseguridad). (2026). Balance de ciberseguridad 2025. https://www.incibe.es

IoT Analytics. (2025). Leading generative AI companies and AI hardware market shares. https://iot-analytics.com

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., ... Barnes, B. (2025). Measuring AI ability to complete long tasks. Model Evaluation and Threat Research.

Lagarde, C. (2025). Speeches on AI and financial system stability. European Central Bank.

Lombardi, M. J., Schrimpf, A., & Sushko, V. (2024, September). The market turbulence and carry trade unwind of August 2024. *BIS Bulletin*, No. 90. Bank for International Settlements. https://www.bis.org/publ/bisbull90.pdf

Microsoft. (2024, July 20). CrowdStrike-related Windows outage update. Microsoft Corporation.

NIST (National Institute of Standards and Technology). (2023). AI Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework

NIST (National Institute of Standards and Technology). (2024). Voluntary pre-deployment evaluation agreements with frontier AI developers. U.S. AI Safety Institute / Center for AI Standards and Innovation.

OCC (Office of the Comptroller of the Currency). (2020, August 6). OCC assesses $80 million civil money penalty against Capital One [Press release]. https://www.occ.gov

OpenAI. (2025). Preparedness Framework, version 2. https://openai.com

OpenAI. (2025, December). BBVA and OpenAI collaborate to transform global banking. https://openai.com/index/bbva-collaboration-expansion/

Parametrix. (2024). CrowdStrike's impact on the Fortune 500: Insurance industry implications. Parametrix Insurance.

Pindrop. (2025). 2025 voice intelligence and security report. Pindrop Security.

Reuters. (2026, April 15). Bank of England's Bailey warns Mythos could "crack the whole cyber risk world open".

Reuters. (2026, May 18). Anthropic to brief Financial Stability Board on cyber flaws exposed by Mythos, FT reports.

SEC (U.S. Securities and Exchange Commission). (2023). Proposed rule: Conflicts of interest associated with the use of predictive data analytics by broker-dealers and investment advisers (Release No. 34-97990) [Withdrawn June 2025].

SEC (U.S. Securities and Exchange Commission). (2024, March 18). SEC charges two investment advisers with making false and misleading statements about their use of artificial intelligence [Press release].

Sophos. (2024). The state of ransomware in financial services 2024. https://www.sophos.com

Sullivan & Cromwell. (2026, April). Federal banking agencies issue revised guidance on model risk management. https://www.sullcrom.com

Sumsub. (2025). Identity fraud report 2025–2026. https://sumsub.com/fraud-report-2025/

Synergy Research Group. (2025). Cloud market share trends: Q1 2025. https://www.srgresearch.com

The Next Web. (2026, May). Anthropic Mythos AI finds thousands of zero-day vulnerabilities as Fed and Treasury convene bank CEOs on cyber risk. https://thenextweb.com

UK AI Security Institute. (2026a). Evaluation of Claude Mythos Preview cyber capabilities. https://www.aisi.gov.uk

UK AI Security Institute. (2026b). Frontier AI trends report. https://www.aisi.gov.uk/frontier-ai-trends-report

U.S. Commodity Futures Trading Commission & U.S. Securities and Exchange Commission. (2010, September 30). Findings regarding the market events of May 6, 2010 [Joint report]. https://www.sec.gov/sec-cftc-prelimreport.pdf

U.S. Department of Justice. (2018, March 14). North Korean regime-backed programmer charged with conspiracy to conduct multiple cyber attacks and intrusions [Press release].

U.S. Department of the Treasury. (2024a, March). Managing artificial intelligence-specific cybersecurity risks in the financial services sector. https://home.treasury.gov

U.S. Department of the Treasury. (2024b). Sanctions and statements related to ICBC ransomware attack.

U.S. Department of the Treasury. (2024c, December). Uses, opportunities, and risks of artificial intelligence in the financial services sector [Report following Request for Information]. https://home.treasury.gov

World Bank. (2023). Bank concentration data: 3-bank asset concentration ratios [Financial Development and Structure Dataset]. World Bank Group.

XBOW. (2026). Mythos for offensive security: XBOW's independent evaluation. https://xbow.com/blog/mythos-offensive-security-xbow-evaluation

Yahoo Finance. (2026, May). Anthropic to share Mythos-linked cyber weakness findings with financial regulator. https://finance.yahoo.com