INVESTIGACIONES DE FUNCAS
40/2026

INTELIGENCIA ARTIFICIAL Y ESTABILIDAD FINANCIERA: LOS RIESGOS BANCARIOS EN LA ERA DE LA IA DE VANGUARDIA — UNA EVALUACIÓN COMPARATIVA DE ESPAÑA, LA UNIÓN EUROPEA Y ESTADOS UNIDOS

Francisco Rodríguez Fernández

Junio 2026

funcas

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

Todos los derechos reservados. Queda prohibida la reproducción total o parcial de esta publicación, así como la edición de su contenido por medio de cualquier proceso reprográfico o fónico, electrónico o mecánico, especialmente imprenta, fotocopia, microfilm, *offset* o mimeógrafo, sin la previa autorización escrita del editor.

ISSN: 3045-8994

INTELIGENCIA ARTIFICIAL Y ESTABILIDAD FINANCIERA: LOS RIESGOS BANCARIOS EN LA ERA DE LA IA DE VANGUARDIA — UNA EVALUACIÓN COMPARATIVA DE ESPAÑA, LA UNIÓN EUROPEA Y ESTADOS UNIDOS

Francisco Rodríguez Fernández

Universidad de Granada y Funcas

Resumen

Esta investigación toma el informe de Anthropic y el FSB de mayo de 2026 sobre el modelo Claude Mythos Preview como punto de referencia empírico para una evaluación comparativa de cómo se posicionan España, la Unión Europea (UE) y Estados Unidos para gestionar los riesgos bancarios relacionados con la IA. Defendemos tres tesis. En primer lugar, la IA de vanguardia ha superado un umbral de capacidad —evidenciado por la tasa de superación del 83,1 % de Mythos en la prueba de referencia de reproducción de vulnerabilidades de CyberGym— que transforma el riesgo cibernético de un peligro operativo idiosincrásico en un riesgo sistémico correlacionado, escalable y, en parte, endógeno. En segundo lugar, la arquitectura regulatoria existente en las tres jurisdicciones es necesaria, pero insuficiente: Estados Unidos excluyó recientemente la IA generativa y agentiva de su guía revisada sobre riesgo de modelos (SR 26-2); la UE cuenta con el marco horizontal más prescriptivo (Ley de IA, DORA), pero se encuentra en una fase temprana de implementación; y España se ha adelantado a sus homólogos de la UE a través de la AESIA, al tiempo que opera con recursos técnicos sustancialmente menores que los de los bancos centrales comparables. En tercer lugar, la dependencia de unos pocos proveedores de modelos de vanguardia y de hiperescaladores en la nube —los tres principales proveedores representan el 44 % de los modelos de IA de terceros utilizados por las entidades financieras del Reino Unido— convierte un problema de organización industrial en uno de estabilidad financiera. Desarrollamos una taxonomía de riesgos en seis dimensiones, presentamos cuatro análisis de escenarios calibrados con puntos de referencia del mundo real (SVB 2023, el *flash crash* de 2010, la liquidación de posiciones de *carry* del yen de agosto de 2024 y la interrupción de CrowdStrike de julio de 2024) y ofrecemos un análisis en profundidad específico para España. Concluimos con siete recomendaciones de política dirigidas a supervisores, autoridades macroprudenciales, el sector y los laboratorios de IA de vanguardia.

Palabras clave: inteligencia artificial, estabilidad financiera, supervisión bancaria, riesgo cibernético, modelos de vanguardia, riesgo sistémico, Ley de IA de la UE, DORA, España.

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

RESUMEN EJECUTIVO

El 18 de mayo de 2026, el Financial Times informó de que Anthropic había acordado informar al Consejo de Estabilidad Financiera (FSB) sobre las vulnerabilidades cibernéticas del sistema financiero mundial identificadas por su modelo Claude Mythos Preview, a petición del gobernador del Banco de Inglaterra y presidente del FSB, Andrew Bailey (Financial Times, 2026; Reuters, 2026). La sesión informativa —precedida por una reunión a puerta cerrada celebrada los días 7 y 8 de abril de 2026 entre el secretario del Tesoro de EE. UU., Scott Bessent, el presidente de la Reserva Federal, Jerome Powell, y los directores ejecutivos de los principales bancos estadounidenses (CNBC, 2026; Sullivan & Cromwell, 2026)— supuso la primera vez que un laboratorio de vanguardia en IA fue tratado como un actor de relevancia sistémica *de facto* por los reguladores financieros mundiales. Bailey, en una intervención en la Universidad de Columbia el 14 de abril de 2026, advirtió de que Mythos podría «abrir de par en par todo el mundo del riesgo cibernético» (Banco de Inglaterra, 2026; Reuters, 2026).

El caso de Mythos es un hecho muy relevante, pero el fenómeno subyacente es estructural. Aproximadamente el 90 % de los bancos importantes de la zona del euro ya utilizan IA (Banco Central Europeo, 2026), el 75 % de las entidades financieras del Reino Unido ha adoptado alguna forma de IA (Breeden, 2024) y los tres principales proveedores externos de IA representan ahora el 44 % de los modelos de IA de terceros utilizados por las instituciones financieras del Reino Unido, frente al 18 % en 2022 (Breeden, 2024). El Consejo de Estabilidad Financiera (2024) ha identificado formalmente cuatro vulnerabilidades relacionadas con la IA —dependencias de terceros, correlaciones de mercado, riesgos cibernéticos y riesgos de modelo— que afectan de manera significativa a la estabilidad financiera. El Comité Científico Asesor del Comité Europeo de Riesgo Sistémico (2025) concluyó que cinco características de la IA —concentración y barreras de entrada, uniformidad de los modelos, dificultades de supervisión, dependencia excesiva y velocidad— podrían amplificar significativamente el riesgo sistémico.

Este documento defiende tres tesis. En primer lugar, la IA de vanguardia ha superado un umbral de capacidad que transforma el riesgo cibernético de un peligro operativo idiosincrásico en un riesgo sistémico correlacionado, escalable y en parte endógeno para el sistema bancario, lo que requiere un tratamiento macroprudencial —y no solo microprudencial—. En segundo lugar, la arquitectura regulatoria existente en EE. UU., la Unión Europea (UE) y España es necesaria pero insuficiente: EE. UU. se basa en cartas de supervisión basadas en principios y recientemente excluyó la IA generativa y agentiva de su guía revisada sobre riesgo de modelo (Junta de la Reserva Federal, FDIC y OCC, 2026); la UE cuenta con el marco horizontal más prescriptivo (Ley de IA de la UE, DORA), pero aún se encuentra en una fase temprana de implementación; y España se ha adelantado a sus homólogos de la UE al crear la primera agencia nacional de supervisión de la IA (AESIA), pero opera con recursos sustancialmente menores que los bancos centrales. En tercer lugar, la dependencia de un pequeño puñado de proveedores de modelos de vanguardia y de hiperescaladores de la nube —AWS con aproximadamente el 30 %, Azure con cerca del 21 %, Google Cloud alrededor del 13 % de la nube global (Synergy Research Group, 2025) y NVIDIA sobre el 92 % de las GPU de centros de datos (IoT Analytics, 2025)— convierte un problema de organización industrial en uno de estabilidad financiera, con un riesgo de concentración que ahora refleja las externalidades de «demasiado grande para quebrar» de los mayores intermediarios financieros.

Cuatro análisis de escenarios, calibrados con puntos de referencia del mundo real (SVB marzo de 2023, caída repentina del 6 de mayo de 2010, liquidación del *carry trade* del yen del 5 de agosto de 2024, interrupción de CrowdStrike del 19 de julio de 2024), sugieren que un evento de retirada de depósitos amplificado por la IAen un G-SIB de tamaño medio podría comprimir aún más la línea temporal de salida de fondos de SVB de 42 000 millones de dólares en un día (Junta de la Reserva Federal, 2023), mientras que un escenario de ataque autónomo de clase Mythos podría reducir el «tiempo de explotación» medio del sector —cuya mediana histórica era de meses— a horas o minutos (Anthropic, 2026a; Anthony Grieco, de Cisco, Anthropic, 2026a). Los bancos españoles, aunque bien capitalizados (CET1 del sector del 13,8 % a junio de 2025; Banco de España, 2025a), se enfrentan a una elevada exposición relativa debido a la alta concentración nacional (CR3 de cerca del 72 %; Banco Mundial, 2023), la condición de G-SIB de Santander y su importante presencia en América Latina y Turquía.

El documento concluye con recomendaciones de política dirigidas a cuatro grupos: supervisores, autoridades macroprudenciales, el sector y los laboratorios de IA de vanguardia. Lo más urgente, según sostiene, es que la consulta sobre «prácticas sólidas» prevista por el FSB (FSB, 2026) formalice tres principios: (i) la designación de proveedores críticos de IA análoga al régimen CTPP de la DORA; (ii) la realización obligatoria de pruebas de «red teaming» previas a la implementación por parte de institutos nacionales independientes (AISI del Reino Unido, CAISI de EE. UU.) para cualquier modelo de vanguardia ofrecido a bancos sistémicos; y (iii) un protocolo coordinado de «divulgación responsable» para las vulnerabilidades detectadas por la IA en la infraestructura financiera, siguiendo el modelo del canal específico Anthropic-FSB que el caso Mythos ha improvisado

1. INTRODUCCIÓN: EL MOMENTO MYTHOS Y EL NEXO ENTRE LA IA Y LAS FINANZAS

1.1. Una intervención reguladora sin precedentes

La secuencia de acontecimientos de abril-mayo de 2026 no tiene precedentes en la historia de la gobernanza financiera internacional posterior a 1944. Anthropic anunció Claude Mythos Preview el 7 de abril de 2026, como parte del «Proyecto Glasswing», describiéndolo como un «modelo de vanguardia de uso general, aún no publicado, que revela un hecho contundente: los modelos de IA han alcanzado un nivel de capacidad de codificación que les permite superar a todos los humanos, salvo a los más expertos, a la hora de encontrar y explotar vulnerabilidades de software» (Anthropic, 2026a). En menos de 24 horas, el secretario del Tesoro, Bessent, y el presidente de la Reserva Federal, Powell, habían convocado a los directores ejecutivos de Bank of America, Citi, Goldman Sachs, Morgan Stanley y Wells Fargo, con Jamie Dimon, de JPMorgan Chase, invitado pero sin poder asistir (CNBC, 2026). Bailey dio a conocer el nombre del modelo públicamente siete días después (Banco de Inglaterra, 2026), el Grupo de Resiliencia Operativa Intermercados del Banco de Inglaterra incluyó a Mythos en su agenda en un plazo de dos semanas (Bloomberg, 2026a), y a mediados de mayo de 2026 la Comisión Australiana de Valores e Inversiones, el Banco Central Europeo, la Agencia de Servicios Financieros de Japón, la Autoridad Monetaria de Singapur y la Comisión de Servicios Financieros de Corea del Sur habían emitido declaraciones públicas o convocado a los directores ejecutivos de los bancos para tratar la amenaza (Reuters, 2026; Nikkei Asia, 2026; NL Times, 2026).

Las capacidades son extraordinarias. Anthropic informa de que Mythos alcanzó una tasa de superación del 83,1 % en la prueba de referencia de reproducción de vulnerabilidades de ciberseguridad de CyberGym, frente al 66,6 % de Claude Opus 4.6 (Anthropic, 2026a). La evaluación independiente de XBOW reveló que Mythos redujo los falsos negativos en la detección de errores en un 42 % en comparación con Opus 4.6 en pruebas de vulnerabilidad web en tiempo real, y en un 55 % cuando se proporcionaba el código fuente (XBOW, 2026). El Instituto de Seguridad de la IA del Reino Unido informó de que Mythos resolvió 22 de los 32 pasos de una simulación de ataque a la red corporativa denominada «Last Ones», que a los expertos humanos les lleva unas 20 horas, y que el tiempo de duplicación de la capacidad de las tareas cibernéticas autónomas se ha reducido de aproximadamente 8 meses en noviembre de 2025 a unos 4,7 meses en febrero de 2026 (Instituto de Seguridad de la IA del Reino Unido, 2026a, 2026b). Mozilla lanzó correcciones para 271 vulnerabilidades en Firefox, 150 detectadas por Mythos en una sola ronda de evaluación (The Next Web, 2026).

1.2. Por qué se trata de un problema de estabilidad financiera, y no solo de ciberseguridad

La opinión ortodoxa en 2024 —expresada en el influyente informe del FSB Implicaciones de la inteligencia artificial para la estabilidad financiera (FSB, 2024)— era que la IA introduce vulnerabilidades operativas y cibernéticas conocidas que pueden abordarse con adaptaciones de los marcos existentes. El momento Mythos pone en entredicho esa opinión de tres maneras.

En primer lugar, la velocidad de la progresión de las capacidades. El director técnico de CrowdStrike, Elia Zaitsev, observó que «el intervalo entre el descubrimiento de una vulnerabilidad y su explotación por parte de un adversario se ha reducido drásticamente: lo que antes llevaba meses, ahora ocurre en minutos con la IA» (Anthropic, 2026a). Cuando el tiempo de explotación se reduce por debajo del tiempo de parcheo, el equilibrio de la ciberdefensa —una carrera probabilística entre defensores y atacantes— se inclina en contra de los operadores tradicionales, cuyos ciclos de parcheo están calibrados para el equilibrio histórico, más lento.

En segundo lugar, la concentración de la capacidad de IA de vanguardia. Aproximadamente 40 organizaciones tienen acceso a Mythos Preview, con 12 socios de lanzamiento nombrados (AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JP MorganChase, la Fundación Linux, Microsoft, NVIDIA, Palo Alto Networks y la propia Anthropic) (Anthropic, 2026a). Anthropic acordó no distribuir el modelo más ampliamente a petición de la Casa Blanca (Financial Times, 2026; Yahoo Finance, 2026). La asimetría es grave: todas las G-SIB se enfrentan al mismo conjunto de amenazas residuales, pero solo una pequeña parte tiene acceso a la capacidad defensiva correspondiente. Como observó Frank Elderson, del Banco Central Europeo, en mayo de 2026, esto «no es una

excusa para la inacción», pero hace que «sea aún más crítico que los bancos den un paso al frente y actúen ahora» (NL Times, 2026).

En tercer lugar, la naturaleza correlacionada de la amenaza. La dependencia del sector financiero de un pequeño número de pilas de *software* comunes (Microsoft Windows, Linux, Chrome/Firefox, marcos Java comunes, SWIFT, Fedwire, T2/TARGET2) significa que una sola clase de exploit del tipo Mythos podría afectar simultáneamente a docenas o cientos de instituciones. La misma lógica de correlación que impulsa el riesgo sistémico en las exposiciones financieras se aplica ahora a las exposiciones operativas y cibernéticas, y el conjunto de herramientas reguladoras para gestionar las primeras está mucho más desarrollado que para las segundas.

1.3. Alcance y hoja de ruta

Este documento sintetiza la evidencia regulatoria, supervisora, académica y del sector desde 2018 hasta mayo de 2026 para elaborar una evaluación comparativa de cómo se posicionan España, la Unión Europea y los Estados Unidos para gestionar el riesgo de estabilidad financiera relacionado con la IA. La sección 2 establece el marco conceptual. La sección 3 desarrolla una taxonomía de los riesgos bancarios relacionados con la IA. La sección 4 analiza las aplicaciones positivas de la IA como contrapeso. La sección 5 compara las arquitecturas regulatorias de las tres jurisdicciones. La sección 6 construye un modelo de canales de contagio. La sección 7 presenta cuatro análisis de escenarios calibrados. La sección 8 examina el perfil de exposición específico de España. La sección 9 establece recomendaciones de política. La sección 10 concluye.

2. MARCO CONCEPTUAL: CÓMO LA IA PROPAGA EL RIESGO EN LA BANCA

2.1. Las tres capas de transmisión

La IA propaga el riesgo en el sistema bancario a través de tres capas anidadas. La capa más interna es la de nivel empresarial: la adopción de la IA por parte de un banco en la calificación crediticia, la detección de fraudes, la negociación o el servicio al cliente introduce riesgos específicos del modelo de error, sesgo y fallo operativo. Este es el ámbito clásico de la gestión microprudencial del riesgo de modelo —SR 11-7 de la Reserva Federal (Junta de Gobernadores del Sistema de la Reserva Federal y OCC, 2011) y su sucesora de 2026 (SR 26-2; Junta de la Reserva Federal, FDIC y OCC, 2026), el marco IRB de aprendizaje automático de la EBA (EBA, 2023) y la guía revisada de modelos internos del BCE (BCE, 2025c).

La capa intermedia es a nivel sectorial: cuando muchos bancos adoptan sistemas de IA similares entrenados con datos similares y arquitecturas similares, los riesgos idiosincrásicos a nivel de empresa se correlacionan. El comportamiento gregario algorítmico, el envenenamiento de datos de fuentes comunes y las respuestas procíclicas sincronizadas a las perturbaciones transforman el riesgo de modelo en riesgo de mercado. El FMI (2024) y el FSB (2024) identifican esto como el principal canal novedoso introducido por la adopción generalizada de la IA.

La capa más externa es a nivel de infraestructura: las dependencias de unos pocos proveedores de modelos de vanguardia, hiperescaladores en la nube y proveedores de hardware especializado (NVIDIA) crean riesgos de concentración que se asemejan a las externalidades de «demasiado grande para quebrar» de los grandes intermediarios financieros, pero con problemas críticos de perímetro regulatorio. NVIDIA, Anthropic, OpenAI, AWS, Microsoft Azure y Google Cloud no se encuentran dentro del perímetro de supervisión directa de ningún regulador bancario, a pesar de que la continuidad operativa de los sistemas bancarios depende cada vez más de ellos.

2.2. Las cinco características amplificadoras de la IA

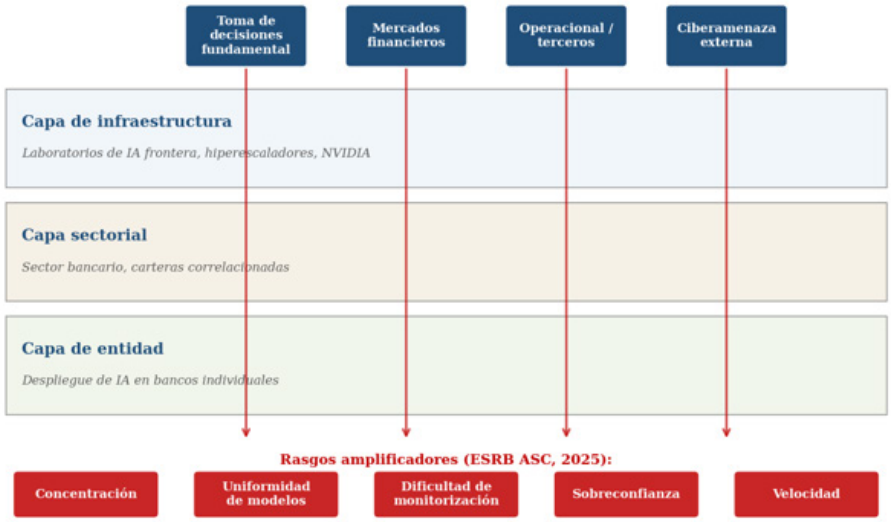
El Comité Científico Asesor del ESRB (2025) ofrece la tipología más rigurosa de por qué la IA es sistémicamente distintiva. De las once características evaluadas, cinco amplifican significativamente las

vulnerabilidades del sistema financiero: concentración y barreras de entrada, uniformidad de los modelos, retos de supervisión, dependencia excesiva y confianza excesiva, y velocidad. Cada una de ellas se corresponde con una o varias de las fuentes clásicas de riesgo sistémico articuladas en la literatura macroprudencial: desajustes de liquidez, exposiciones comunes, interconexión, falta de sustituibilidad y apalancamiento.

El Comité de Política Financiera del Banco de Inglaterra (Banco de Inglaterra, 2025) identifica cuatro canales a través de los cuales los riesgos relacionados con la IA se transmiten a la estabilidad financiera: (i) mayor uso de la IA en la toma de decisiones financieras fundamentales de los bancos; (ii) mayor uso de la IA en los mercados financieros (negociación correlacionada, comportamiento gregario); (iii) riesgos operativos derivados de los proveedores de servicios de IA; y (iv) un entorno cambiante de amenazas cibernéticas externas. Estos cuatro canales pueden relacionarse con las cinco características de la JERS y las cuatro vulnerabilidades del FSB para elaborar un modelo conceptual integrado, que se resume en la figura 1.

Figura 1. MAPA CONCEPTUAL INTEGRADO DE TRANSMISIÓN DEL RIESGO DE IA A LA ESTABILIDAD BANCARIA

(Tres capas verticales x Cuatro canales de transmisión x Cinco rasgos amplificadores)



Notas: (Diagrama conceptual, sin escala).
Tres capas verticales —empresa, sector, infraestructura— se cruzan con cuatro canales horizontales —toma de decisiones fundamentales, mercados, operativo/terceros, ciberamenazas externas—. Las cinco características amplificadoras del ESRB (concentración, uniformidad, dificultades de supervisión, dependencia excesiva, velocidad) actúan como modificadores transversales que magnifican la magnitud y la correlación de las perturbaciones transmitidas a través de cualquier combinación de canal y capa.

2.3. Riesgo endógeno frente a riesgo exógeno

Una característica distintiva del riesgo de IA es que se sitúa incómodamente entre las categorías estándar de riesgo endógeno y exógeno en la teoría macroprudencial. Las herramientas de IA desplegadas por los propios bancos contribuyen de forma endógena a la potencial prociclicidad, el comportamiento gregario y las retiradas masivas correlacionadas. Las herramientas de IA desplegadas por adversarios —ya sean actores estatales, ciberdelincuentes o activistas— generan choques exógenos que el sistema financiero debe absorber. Los modelos de vanguardia como Mythos se sitúan simultáneamente a ambos lados de esta división: la misma capacidad que permite a JPMorganChase o Microsoft reforzar sus sistemas también representa una amenaza latente si, llega a manos de actores menos alineados. Esta dualidad, el problema del doble uso, es el núcleo empírico del asunto Mythos y la novedad regulatoria más acuciante del momento.

3. TAXONOMÍA DE LOS RIESGOS BANCARIOS RELACIONADOS CON LA IA

En esta sección se desarrolla una taxonomía de seis dimensiones. La tabla 1 resume la estructura; las seis subsecciones siguientes la profundizan.

TABLA 1. TAXONOMÍA DE LOS RIESGOS BANCARIOS RELACIONADOS CON LA IA

Dimensión	Principales vectores de riesgo	Instrumentos reguladores clave	Referencia empírica
Operacional y cibernético	Doble uso del modelo de frontera; concentración de terceros en proveedores de IA; deficiencias en la gestión del riesgo de modelo	SR 26-2 (EE. UU.); DORA, arts. 28-44 (UE); Ley de IA de la UE, art. 15; Marco de pruebas de resistencia cibernéticas del BCE	Mythos (abril de 2026); *ransomware* del ICBC (noviembre de 2023); interrupción del servicio de CrowdStrike (julio de 2024)
Sistémico/ macroprudencial	Efecto manada algorítmico; prociclicidad; eventos repentinos impulsados por la IA; retiradas masivas de depósitos amplificadas por la IA	Informe n.º 16 del Comité de Supervisión del Sistema Europeo (ESRB) (2025); Informe sobre la Estabilidad Financiera Global (GFSR) del FMI, octubre de 2024, cap. 3; Informe sobre la Estabilidad Financiera (FSiF) del Banco de Inglaterra, abril de 2025	Quiebra de SVB (marzo de 2023); caída repentina del 6 de mayo de 2010; liquidación de posiciones de *carry trade* en yenes del 5 de agosto de 2024
Crédito y mercado	Sesgo en la puntuación crediticia de la IA; negociación algorítmica; fraude a gran escala mediante *deepfakes*	Anexo III(5)(b) de la Ley de IA de la UE; Circular 2023-03 de la CFPB; RTS 6 de la MiFID II	Fraude con *deepfakes* de Arup por valor de 200 millones de HK$ (enero de 2024)
Gobernanza, riesgo de modelo, explicabilidad	Riesgo de caja negra; alucinaciones en la IA orientada al cliente; lagunas en la validación de modelos	SR 26-2 (EE. UU.); Seguimiento del ML-IRB de la EBA (2023); Artículos 13-15 de la Ley de IA de la UE	Conclusión de la ESMA: ChatGPT alucina en aproximadamente el 19,5 % de las respuestas (2024)
Blanqueo de capitales, fraude, conducta	Identidad sintética; clonación de voz; fraude de *deepfak* del director general; desinformación generada por IA	Informe cibernético sobre IA del Tesoro (marzo de 2024); informes del IC3 del FBI; directrices del GAFI	Pérdidas de 16 600 millones de dólares según el IC3 del FBI en 2024; aumento de los *deepfakes* en un 1300 % según Pindrop (2025)
Concentración/oligopolio	Dependencia de Anthropic, OpenAI, Google DeepMind; dominio de NVIDIA; triopolio de AWS/Azure/GCP	Régimen DORA CTPP; recomendaciones del FSB de 2024	El 44 % de los modelos de IA del sector financiero del Reino Unido proceden de los tres principales proveedores; NVIDIA, aproximadamente el 92 % de las GPU de los centros de datos

Fuente: Elaboración propia a partir de la normativa vigente.

3.1. Riesgo operativo y cibernético: el problema del doble uso al estilo Mythos

El *Informe sobre Estabilidad Financiera Mundial* del FMI de 2024 estimó que «las pérdidas cibernéticas que una empresa financiera podría sufrir una vez cada década podrían alcanzar los 2.200 millones de dólares, frente a los aproximadamente 300 millones de dólares de 2017» (FMI, 2026). La referencia histórica es aleccionadora: el robo de SWIFT al Banco de Bangladesh (febrero de 2016) supuso un coste de aproximadamente 81 millones de dólares de los 951 millones que se intentaron sustraer, de los cuales solo se recuperaron 15 millones (BAE Systems, 2016; Departamento de Justicia de EE. UU., 2018); la filtración

de Capital One (julio de 2019) expuso 106 millones de registros de clientes y supuso una multa civil de 80 millones de dólares impuesta por la OCC, además de un acuerdo de demanda colectiva de 190 millones de dólares (OCC, 2020; Administración del Acuerdo de Capital One , 2022); el ataque de *ransomware* al ICBC (8-9 de noviembre de 2023) interrumpió la liquidación de más de 9.000 millones de dólares en activos respaldados por el Tesoro (Departamento del Tesoro de EE. UU., 2024b); y la campaña de día cero de MOVEit (mayo-junio de 2023) afectó a más de 2500 organizaciones y 66 millones de personas, incluidas numerosas víctimas del sector financiero (CISA y FBI, 2023; Emsisoft, 2023).

La introducción de modelos de vanguardia cambia radicalmente la superficie de ataque. El informe del Tesoro de EE. UU. titulado «Gestión de los riesgos de ciberseguridad específicos de la inteligencia artificial» (Departamento del Tesoro de EE. UU., 2024a) —publicado en virtud de la Orden Ejecutiva 14110— identificó cuatro vulnerabilidades específicas de la IA (envenenamiento de datos, fuga de datos durante la inferencia, evasión y extracción de modelos) y advirtió de que «la IA permite a los actores maliciosos suplantar la identidad de personas, como empleados y clientes de instituciones financieras, de formas que antes eran mucho más difíciles». El informe de marzo de 2024 también señaló la creciente brecha de capacidades entre las instituciones grandes y pequeñas como una vulnerabilidad a nivel sectorial.

La gestión del riesgo de los modelos ha sido la principal respuesta microprudencial. La Reserva Federal de EE. UU., la FDIC y la OCC publicaron conjuntamente la SR 26-2 el 17 de abril de 2026, que sustituye a la SR 11-7 (Junta de la Reserva Federal, FDIC y OCC, 2026). La guía revisada tiene aproximadamente la mitad de extensión que su predecesora de 2011 y elimina el requisito de revalidación anual. Es fundamental destacar que la IA generativa y la IA agentiva quedan explícitamente fuera de su ámbito de aplicación (SR 26-2, nota al pie 3). El vicepresidente Bowman defendió esta decisión el 1 de mayo de 2026: «La guía revisada se aplica ahora de forma restringida a los modelos tradicionales y a las aplicaciones básicas de IA» (Bowman, 2026). El enfoque de la UE es más prescriptivo: los sistemas de IA de alto riesgo contemplados en el artículo 6 de la Ley de IA deben cumplir con los artículos 9 (gestión de riesgos), 10 (gobernanza de datos), 11 (documentación técnica), 13 (transparencia), 14 (supervisión humana) y 15 (precisión, solidez y ciberseguridad) (Reglamento (UE) 2024/1689). La Guía revisada del BCE sobre modelos internos (BCE, 2025c) introduce, por primera vez, las expectativas del BCE respecto a los modelos de aprendizaje automático en el IRB.

Podría decirse que el aspecto de la concentración de terceros es ahora más importante que el del riesgo de modelo. La encuesta sobre IA de 2024 del Banco de Inglaterra reveló que el 44 % de los modelos de IA de terceros utilizados por las entidades financieras del Reino Unido proceden de los tres principales proveedores, frente al 18 % en 2022 (Breeden, 2024). La ronda de financiación de serie G de Anthropic, de 30 000 millones de dólares, en febrero de 2026, con una valoración post-money de 380 000 millones de dólares (Anthropic, 2026b), y la valoración implícita de OpenAI, de aproximadamente 852 000 millones de dólares (Bloomberg, 2026; citada en foreignpolicyjournal.com, 2026) hacen que estos proveedores sean más valiosos que todos los bancos, salvo los más grandes, al tiempo que operan por completo fuera del perímetro prudencial. El compromiso de Amazon de proporcionar recursos informáticos a Anthropic, por un total de entre 25 000 y 33 000 millones de dólares (Anthropic, 2025), y la participación de Microsoft en OpenAI, valorada en 13 000 millones de dólares (CNBC, 2025), entrelazan aún más la IA de vanguardia con la concentración en la nube.

3.2. Riesgo sistémico y macroprudencial

Cabe destacar tres canales macroprudenciales. En primer lugar, el comportamiento gregario algorítmico y la prociclicidad: el FMI (2024) informó de que el comercio algorítmico ya representa aproximadamente el 70 % de la renta variable estadounidense y aproximadamente el 50 % del volumen de futuros. La autoridad neerlandesa del mercado energético constató que «el aprendizaje automático se utiliza de forma implícita o explícita en entre el 80 % y el 100 % de los algoritmos [de las empresas de negociación]» (FMI, 2024). Según informó Coalition Greenwich (2024), el 37 % del flujo de órdenes de compra de acciones estadounidenses se ejecutó mediante algoritmos o enrutadores inteligentes de órdenes en 2023, frente al 35 % del año anterior. Cuando los modelos comparten arquitecturas, datos o métodos de entrenamiento, sus respuestas a las perturbaciones se correlacionan, lo que puede ser benigno en mercados tranquilos y catastrófico en mercados

bajo presión. El FMI (2024) constató que los ETF impulsados por IA experimentaron una mayor rotación que los ETF activos o pasivos, con un pico de rotación en marzo de 2020 —una prueba conductual de prociclicidad—.

En segundo lugar, las retiradas masivas de depósitos amplificadas por la IA. La quiebra del Silicon Valley Bank entre el 9 y el 10 de marzo de 2023 sentó las bases empíricas para comprender las retiradas masivas impulsadas por las redes sociales. El informe de la Junta de la Reserva Federal (2023) documenta que «las salidas de depósitos superaron los 40.000 millones de dólares el 9 de marzo, y la dirección esperaba 100.000 millones más al día siguiente», lo que implica un total de intentos de salida de aproximadamente 142 000 millones de dólares frente a una base de depósitos de 166.000 millones (aproximadamente el 85 %) comprimida en 36 horas. Cookson, Fox, Gil-Bazo, Imbet y Schiller (2025) descubrieron que entre el 8 y el 13 de marzo de 2023 se publicaron 6528 tuits sobre la «corrida» de SVB, y que los bancos situados en el tercil superior de exposición previa en Twitter perdieron 6,6 puntos porcentuales más de valor bursátil durante el periodo de la corrida —un efecto comparable en magnitud a una desviación estándar de los depósitos no asegurados. La implicación es que el contenido generado por IA —redes coordinadas de bots, declaraciones «filtradas» de ejecutivos mediante *deepfakes*, publicaciones de «pánico» generadas por IA— podría amplificar una estampida social al estilo de la de 2023 en un orden de magnitud.

En tercer lugar, los eventos repentinos impulsados por la IA. La caída repentina del 6 de mayo de 2010 provocó que el Dow Jones Industrial Average cayera 998,5 puntos (aproximadamente un 9 %) en el transcurso de unos 36 minutos, desencadenada por un programa algorítmico de venta de E-Mini por valor de 4100 millones de dólares (CFTC y SEC de EE. UU., 2010). La liquidación de las operaciones de *carry trade* con el yen del 5 de agosto de 2024 provocó que el índice TOPIX perdiera un 12 % en un solo día y que el VIX superara brevemente niveles no vistos desde la COVID-19, con el S&P 500 cayendo un 3,0 % ese mismo día (Lombardi et al., 2024). Estos episodios se produjeron sin la participación de la IA de vanguardia. Futuros eventos repentinos que impliquen estrategias de negociación de IA correlacionadas —especialmente si se trata de un « » que responda a señales o noticias generadas por IA— podrían combinar la velocidad de la caída de 2010 con la magnitud de la liquidación de posiciones de *carry* de 2024.

3.3. Riesgo de crédito y de mercado

La IA para la evaluación de la solvencia crediticia se clasifica explícitamente como un sistema de IA de alto riesgo con arreglo al anexo III, apartado 5, letra b), de la Ley de IA de la UE (Reglamento (UE) 2024/1689), con la única excepción de los sistemas utilizados para la detección del fraude financiero. Esto da lugar al conjunto completo de obligaciones recogidas en los artículos 9 a 16. En EE. UU., la Circular 2023-03 de la CFPB (Oficina para la Protección Financiera del Consumidor, 2023) exige que las notificaciones de medidas adversas especifiquen los motivos reales de la denegación de crédito, incluso cuando dichas denegaciones sean resultado de algoritmos complejos: tal y como declaró el entonces director Chopra, «los acreedores no pueden basarse en la lista de motivos proporcionada en los formularios de muestra… si dichos motivos no indican de forma específica y precisa la(s) razón(es) principal(es) de la acción adversa». El informe «Chatbot Spotlight» de la CFPB (2023) estimó que aproximadamente el 37 % de la población estadounidense interactuó con un *chatbot* bancario en 2022, y se prevé que esta cifra aumente hasta superar el 40 % en 2026.

El *trading* algorítmico está regulado en la UE a través de la RTS 6 de la MiFID II (Reglamento Delegado (UE) 2017/589 de la Comisión) y reforzado por una nota informativa de la AEVM de 26 de febrero de 2026 en la que se señala explícitamente que los sistemas de *trading* algorítmico que se ajusten a la definición de «sistema de IA» de la Ley de IA deben cumplir ambos regímenes. En EE. UU., la propuesta de norma de la SEC de 2023 sobre análisis predictivo de datos (SEC, 2023) fue retirada en junio de 2025; las primeras medidas coercitivas de la SEC contra el «AI-washing» dirigidas a Delphia y Global Predictions (SEC, 2024) —que dieron lugar a sanciones de 225.000 y 175.000 dólares, respectivamente— representan una alternativa más limitada, pero que sienta precedente.

El fraude basado en *deepfakes* se ha extendido rápidamente. El caso Arup (enero de 2024) implicó 15 transferencias fraudulentas por un total de 200 millones de dólares de Hong Kong (aproximadamente 25,6 millones de dólares estadounidenses) en un solo día, inducidas por una videoconferencia *deepfake* en la que aparecían imágenes generadas por IA del director financiero y otros altos cargos (CNN, 2024; Policía de Hong Kong, 2024). Sumsub (2025) informó de un aumento del 180 % en los fraudes sofisticados en 2025 con respecto a 2024, y Pindrop (2025) documentó un aumento del 1.300 % en los intentos de fraude mediante *deepfakes* en 2024, pasando de uno al mes a siete al día de media. El FBI IC3 (2025) registró 16.600 millones de dólares en pérdidas totales notificadas relacionadas con la ciberdelincuencia en 2024, lo que supone un aumento interanual del 33 %, de los cuales 2.770 millones de dólares se atribuyeron al fraude por suplantación de identidad en el correo electrónico empresarial.

3.4. Gobernanza, riesgo de modelo y explicabilidad

El problema de la «caja negra» se ve agravado, y no resuelto, por la última generación de modelos. La declaración pública de la ESMA de 2024 sobre la IA en los servicios de inversión minorista (ESMA, 2024) citaba una investigación empírica que revelaba que ChatGPT alucina en aproximadamente el 19,5 % de las respuestas, una tasa que es inaceptable desde el punto de vista operativo para las aplicaciones de asesoramiento financiero dirigidas al cliente. El Banco de Inglaterra (2024) constató que «solo un tercio de los encuestados afirma tener una comprensión completa de las tecnologías de IA que han implementado» y que el 55 % de todos los casos de uso de IA notificados incluyen algún tipo de toma de decisiones automatizada, aunque solo el 2 % es totalmente autónomo a fecha de abril de 2025 (Banco de Inglaterra, 2025).

Los marcos de gobernanza difieren sustancialmente entre jurisdicciones. El Marco de Gestión de Riesgos de IA del NIST (NIST, 2023) es voluntario, tecnológicamente neutral e intersectorial, y se organiza en torno a cuatro funciones básicas: GOVERN, MAP, MEASURE, MANAGE. La norma NIST AI 600-1, el Perfil de IA Generativa (Autio et al., 2024), añade 12 categorías de riesgo específicas de la IA generativa, entre las que se incluyen la confabulación, el sesgo perjudicial y la integración en la cadena de valor, pero no constituye un perfil específico para los servicios financieros. El capítulo III, sección 2, de la Ley de IA de la UE exige un sistema específico de gestión de riesgos (art. 9), gobernanza de datos (art. 10) y supervisión humana explícita (art. 14) para los sistemas de alto riesgo. El dictamen de la EIOPA de 6 de agosto de 2025 (EIOPA, 2025) aplica principios de gobernanza similares a los seguros.

3.5. Blanqueo de capitales, fraude y riesgo de conducta

El fraude facilitado por la IA generativa ya no es hipotético. La ABE y el BCE informaron conjuntamente de 4.200 millones de euros en fraude de pagos en todo el Espacio Económico Europeo en 2024, y la manipulación de los pagadores (ingeniería social) representó más de la mitad de las transferencias de crédito fraudulentas (ABE y BCE, 2025). El INCIBE español gestionó 122.223 incidentes de ciberseguridad en 2025, un 26 % más que el año anterior, de los cuales el 34 % de los incidentes que afectaron a los operadores esenciales e importantes (ámbito de aplicación de la NIS2) se produjeron en el sector bancario (INCIBE, 2026). A nivel mundial, el FBI IC3 informó de pérdidas por ciberfraude por valor de 16.600 millones de dólares en 2024 (FBI IC3, 2025), y Sophos (2024) reveló que el 65 % de las organizaciones de servicios financieros se vieron afectadas por *ransomware* en 2024, con un pago medio de rescate de 2 millones de dólares para quienes pagaron.

El Grupo de Trabajo sobre Riesgos de IA del FS-ISAC publicó seis libros blancos en febrero de 2024 que abarcaban la IA adversaria, la integración de la IA defensiva, la evaluación de proveedores de IA generativa y las políticas de uso aceptable (FS-ISAC, 2024). Estos marcos sectoriales complementan, pero no sustituyen, a la acción reguladora.

3.6. Riesgo de concentración y oligopolio

La concentración estructural de la IA de vanguardia está ahora bien documentada. Synergy Research Group (2025) informa de que los tres grandes hiperescaladores (AWS con aproximadamente el 30 %, Azure cerca del 21 % y Google Cloud alrededor del 13 %) acaparan el 63 % del gasto empresarial global en la nube, con un gasto en la nube para todo el año 2024 que alcanza aproximadamente los 330.000 millones de dólares. IoT Analytics (2025) estima que NVIDIA poseía cerca del 92 % del mercado de GPU para centros de datos en 2024, en un mercado valorado en 125.000 millones de dólares. En los modelos de vanguardia, el universo de proveedores «relevantes para ASL-4» es aún más reducido: Anthropic, OpenAI, Google DeepMind, con actores secundarios como xAI, Meta AI y Mistral.

En el caso concreto de la banca, Breeden (2024) informó de que el 44 % de los modelos de IA de terceros utilizados por las entidades financieras del Reino Unido proceden de los tres principales proveedores, frente al 18 % en 2022. El informe del Tesoro de diciembre de 2024 sobre la IA en los servicios financieros (Departamento del Tesoro de EE. UU., 2024c) señaló el «riesgo de concentración de terceros (pocas empresas que dominan el mercado de modelos avanzados)» como una de las principales preocupaciones. El régimen de proveedores de TIC críticos de terceros (CTPP) de la DORA (arts. 31-44; Reglamento (UE) 2022/2554), complementado por el Reglamento Delegado (UE) 2024/1502 de la Comisión, de 22 de febrero de 2024, permite a las Autoridades Europeas de Supervisión designar a proveedores terceros críticos e imponer multas de hasta el 1 % del volumen de negocios medio diario a nivel mundial por incumplimiento de las recomendaciones. A mediados de 2026, las designaciones de CTPP aún no se han concretado.

4. EFECTOS POSITIVOS DE LA IA EN LA BANCA: EL CONTRAPESO

Una evaluación equilibrada debe tener en cuenta que la IA también hace que el sistema financiero sea más resiliente en múltiples dimensiones. El BIS (2024) informó de que aproximadamente el 70 % de las empresas de servicios financieros utilizan la IA para mejorar las predicciones de flujo de caja, optimizar la gestión de la liquidez, ajustar la calificación crediticia y mejorar la detección de fraudes. El Tesoro de EE. UU. (comunicado de prensa del 17 de octubre de 2024, citado en Bowman, 2024) atribuyó a las herramientas de detección de fraudes basadas en IA y aprendizaje automático la prevención y recuperación de más de 4.000 millones de dólares en fraudes durante el año fiscal 2024, incluidos aproximadamente 1.000 millones de dólares en fraudes con cheques del Tesoro.

En la concesión de créditos, la IA permite el uso de datos alternativos para conceder crédito a consumidores con historial crediticio escaso, aunque tanto la CFPB como la Ley de IA de la UE insisten en la transparencia de las medidas adversas. En materia de lucha contra el blanqueo de capitales, el Proyecto Aurora del BIS Innovation Hub e iniciativas similares demuestran que la detección de anomalías basada en el aprendizaje automático puede superar a los sistemas basados en reglas. En el servicio de atención al cliente, Erica, de Bank of America, alcanzó los 2.500 millones de interacciones acumuladas a finales de 2024, con 676 millones solo en 2024, atendiendo a 20 millones de usuarios únicos (Bank of America, 2025). BBVA informó de que su implementación de OpenAI ahorró aproximadamente tres horas a la semana por empleado en tareas rutinarias de « », con una participación diaria superior al 80 %, lo que condujo a una ampliación de 3.300 a 11.000 licencias en 2025, y a la cobertura de los 120.000 empleados en 25 países en diciembre de 2025 (OpenAI, 2025; BBVA, 2025).

En materia de modelización de riesgos, el informe de seguimiento de la ABE sobre el aprendizaje automático para los modelos IRB (ABE, 2023) es cautelosamente permisivo, permitiendo el uso del aprendizaje automático en el desarrollo de modelos de probabilidad de impago (PD) siempre que se respeten las restricciones de explicabilidad y estabilidad. La Guía de Modelos Internos revisada del BCE (BCE, 2025c) abre igualmente una vía regulada. El Banco de España está contratando a 25 profesionales de IA y ciencia de datos para desarrollar herramientas de IA para la supervisión (Banco de España, 2025b). El BCE ha implementado herramientas de IA para la supervisión, entre las que se incluyen Athena (búsqueda de documentos), Virtual Lab, Delphi (detección temprana de riesgos), Medusa y Heimdall (evaluaciones de idoneidad) (Supervisión Bancaria del BCE, 2025).

El punto crucial es que el mismo avance en las capacidades que crea un riesgo ofensivo de doble uso (el hallazgo de Mythos de 271 vulnerabilidades en Firefox) también genera un beneficio defensivo (la corrección de esas mismas vulnerabilidades). La cuestión es si el efecto neto sobre la estabilidad financiera es positivo, y la respuesta depende fundamentalmente del marco normativo y de supervisión que rige la divulgación, el acceso y el despliegue.

5. PANORAMA REGULATORIO COMPARATIVO: ESPAÑA, LA UE Y LOS ESTADOS UNIDOS

5.1. Estados Unidos: basado en principios, cada vez más fragmentado

El marco estadounidense es complejo y, en parte, contradictorio en 2026. La Orden Ejecutiva 14110 (30 de octubre de 2023) ordenó a más de 50 entidades federales que llevaran a cabo más de 100 acciones, entre ellas el informe de ciberseguridad de la IA del Tesoro de marzo de 2024 (Departamento del Tesoro de EE. UU., 2024a) y el perfil GenAI 600-1 del NIST (Autio et al., 2024). La Orden Ejecutiva 14110 fue derogada por la Orden Ejecutiva 14148 el 20 de enero de 2025, y la Orden Ejecutiva 14179, de 23 de enero de 2025, declaró que la política de EE. UU. consistía en «mantener y reforzar el dominio mundial de Estados Unidos en materia de IA» (Orden Ejecutiva 14179, 2025). Otra Orden Ejecutiva del 11 de diciembre de 2025 («Eliminación de la obstrucción de la política nacional de inteligencia artificial por parte de las leyes estatales») tiene por objeto prevalecer sobre las leyes estatales en materia de IA.

En el ámbito bancario, el principal instrumento microprudencial es la norma SR 26-2 (Junta de la Reserva Federal et al., 2026), que sustituye a la norma SR 11-7 (Junta de Gobernadores del Sistema de la Reserva Federal y OCC, 2011) y a la norma SR 21-8. Tal y como se expone en la sección 3.1, la IA generativa y la IA con capacidad de agencia quedan expresamente excluidas de su ámbito de aplicación. El riesgo de terceros se rige por la SR 23-4 (la Guía Interinstitucional de 2023 sobre Relaciones con Terceros). El informe sobre *chatbots* de la CFPB y la Circular 2023-03 regulan la IA dirigida al consumidor. La SEC ha utilizado las facultades antifraude existentes para perseguir el «AI washing» (SEC, 2024). El Centro de Estándares de IA e Innovación en la Industria (CAISI) del NIST, renombrado a partir de la US AISI en junio de 2025, ha firmado acuerdos voluntarios de evaluación previa a la implementación con Anthropic, OpenAI, Google, Microsoft y xAI (NIST, 2024).

La medida de supervisión más trascendental del periodo estudiado fue la reunión celebrada los días 7 y 8 de abril de 2026 entre el Tesoro, la Reserva Federal y los directores ejecutivos de los bancos sobre Mythos, tal y como confirmó la vicepresidenta Bowman en su discurso del 1 de mayo de 2026 (Bowman, 2026). Bowman también reveló que preside el Comité Permanente de Cooperación Supervisora y Regulatoria del FSB y que el informe de consulta sobre prácticas sólidas de IA se publicaría para comentarios en el tercer trimestre de 2026.

5.2. La Unión Europea: el marco horizontal más prescriptivo

El marco de la UE se sustenta en tres pilares. En primer lugar, la Ley de IA de la UE (Reglamento (UE) 2024/1689) entró en vigor el 1 de agosto de 2024, con una aplicación escalonada: prohibiciones y obligaciones de alfabetización en IA a partir del 2 de febrero de 2025; las normas, la gobernanza y las sanciones relativas a la IA de propósito general (GPAI) a partir del 2 de agosto de 2025; las principales obligaciones de alto riesgo a partir del 2 de agosto de 2026; y las normas sobre componentes de seguridad del artículo 6, apartado 1, a partir del 2 de agosto de 2027. La IA para la puntuación crediticia se considera explícitamente de alto riesgo en virtud del anexo III, apartado 5, letra b). Los modelos de IA de uso general entrenados con una capacidad de cálculo acumulada superior a 10^{25} FLOP se clasifican como de riesgo sistémico (art. 51). Las sanciones ascienden a 35 millones de euros o al 7 % de la facturación anual mundial por las prohibiciones del artículo 5, y a 15 millones de euros o al 3 % por otras obligaciones.

En segundo lugar, el Reglamento DORA (Reglamento (UE) 2022/2554), aplicable a partir del 17 de enero de 2025, constituye la columna vertebral de la resiliencia operativa. Sus cinco pilares —gestión de riesgos de

TIC, gestión y notificación de incidentes, pruebas de resiliencia operativa digital, incluida la TLPT cada tres años (artículos 26-27), riesgos de terceros en materia de TIC, incluida la designación de CTPP (artículos 31-44), e intercambio de información— se aplican a aproximadamente 20 categorías de entidades financieras, además de a los proveedores externos de TIC. Las normas técnicas de regulación (RTS) sobre la TLPT se publicaron como Reglamento Delegado (UE) 2025/1190 de la Comisión el 18 de junio de 2025. El primer ciclo DORA-TLPT para las entidades designadas debe completarse antes del 17 de enero de 2028.

En tercer lugar, el marco de supervisión del MUS. Las Prioridades de Supervisión del MUS del BCE para 2026-2028 (BCE, 2025d) señalan que los riesgos operativos y de TIC siguen recibiendo las peores puntuaciones medias en el SREP. La prueba de resistencia de ciberresiliencia del BCE de 2024 abarcó 109 bancos supervisados directamente, de los cuales 28 fueron sometidos a pruebas exhaustivas que incluían la recuperación real de los sistemas informáticos (BCE, 2024b). El Boletín Macroprudencial del BCE de febrero de 2025 (BCE, 2025a) amplió el análisis a perspectivas macroprudenciales, identificando tres canales de contagio para los incidentes cibernéticos: operativo, financiero y de confianza.

El Informe n.º 16 del Comité Científico Asesor de la JERS (JERS, 2025), publicado el 4 de diciembre de 2025, ofrece el análisis más riguroso del riesgo sistémico europeo. El boletín de Frank Elderson del 13 de mayo de 2026 sobre Mythos (NL Times, 2026) y el insistente énfasis de Christine Lagarde en el riesgo de la IA (por ejemplo, Lagarde, 2025) demuestran la gran relevancia política de la IA dentro del BCE.

5.3. España: La primera agencia nacional de supervisión de la IA en la UE

España ha seguido una estrategia de adelantarse a las obligaciones de la UE. La AESIA (Agencia Española de Supervisión de Inteligencia Artificial) fue creada por el Real Decreto 729/2023, de 22 de agosto de 2023 (Boletín Oficial del Estado, 2023), lo que convirtió a España en el primer Estado miembro de la UE en establecer una agencia nacional de supervisión de la IA, once meses antes de que entrara en vigor la Ley de IA de la UE. La autoridad de inspección de las prácticas prohibidas en materia de IA comenzó a ejercer sus funciones el 2 de febrero de 2025, con plenas facultades sancionadoras a partir del 2 de agosto de 2025. A finales de 2025, la plantilla ascendía a aproximadamente 30 personas.

El Banco de España mantiene la responsabilidad de la supervisión bancaria (dentro del MUS) y publica su *Informe de Estabilidad Financiera* semestralmente. El IEF de otoño de 2025 (Banco de España, 2025a) señala que el sector bancario español presenta una rentabilidad sobre recursos propios (ROE) del 14,6 % y un capital básico de nivel 1 (CET1) del 13,8 % a junio de 2025, y que los riesgos cibernéticos e híbridos se tratan en la sección de riesgos geopolíticos. El banco ha puesto en marcha un grupo de trabajo transversal sobre IA, al tiempo que señala que el Banco de España solo cuenta con el 5 % de los recursos tecnológicos de sus homólogos más grandes del Eurosistema, a pesar de poseer el 12 % del capital del BCE.

La CNMV abordó la IA en los mercados de capitales en su Plan de Actividades 2024 (CNMV, 2024) y ha contratado a 76 nuevos empleados para la supervisión de MiCA y DORA. Sus actividades en materia de IA incluyen la exploración de la IA generativa en los servicios de inversión, el análisis de entornos de pruebas (*sandbox*) y la mejora de los procesos internos.

El marco normativo más amplio de España incluye la actualización de 2024 de la ENIA (Estrategia de Inteligencia Artificial), con una financiación de 1.500 millones de euros para el periodo 2024-2025 (además de los 600 millones de euros del periodo 2021-2023), y una cartera de iniciativas que incluye ALIA (modelo generativo abierto en español), Quantum Spain e IA en Cadenas de Valor (Gobierno de España, 2024).

TABLA 2. MATRIZ COMPARATIVA NORMATIVA: ESPAÑA, UE, EE. UU.

Dimensión	España	UE (MSU, Ley de IA, DORA)	EE. UU. (Fed, OCC, FDIC, CFPB, SEC, NIST)
Legislación horizontal sobre IA	Ley de IA de la UE (efecto directo); AESIA como supervisor	Ley de IA de la UE 2024/1689 (en vigor el 1 de agosto de 2024)	Ninguna; el Decreto Ejecutivo 14179 se decanta por la desregulación
Específica para la banca en materia de IA	SR 26-2 no aplicable; se aplica la Ley de IA	Ley de IA de la UE, anexo III, apartado 5, letra b): puntuación crediticia de alto riesgo; guía del BCE sobre modelos internos (2025)	SR 26-2 (abril de 2026): excluye la IA general y la IA agentiva
Resiliencia operativa	DORA por efecto directo	DORA (Reg. 2022/2554): se aplica a partir del 17 de enero de 2025	SR 23-4 (orientaciones de terceros de 2023); Manual de TI de la FFIEC (voluntario)
Pruebas de estrés cibernético	TIBER-EU/TIBER-ES; participación del CBE	TLPT cada 3 años (arts. 26-27) para 2028; prueba de resistencia cibernética del BCE con 109 bancos en 2024	Equivalente a CBEST a nivel estatal (NYDFS); no hay pruebas de estrés cibernético federales horizontales
Normas GPAI / modelo de frontera	Efecto directo de la Ley de IA de la UE	Art. 51 GPAI de riesgo sistémico (10^{25} FLOP); Código de prácticas de la GPAI (julio de 2025)	Memorandos de entendimiento voluntarios de la CAISI con Anthropic, OpenAI, Google, Microsoft y xAI
Sanciones	Según la Ley de IA de la UE	Hasta 35 millones de euros / 7 % de la facturación mundial	SEC, OCC, CFPB, FRB *ad hoc*; sin régimen específico para la IA
Agencia nacional de IA	AESIA (desde agosto de 2023; operativa en febrero de 2025)	Oficina Europea de IA (desde enero de 2024)	NIST CAISI (renombrada a partir de la AISI de EE. UU. en junio de 2025)

6. CANALES DE CONTAGIO Y AMPLIFICACIÓN SISTÉMICA

Un modelo de cómo se propagan las perturbaciones relacionadas con la IA a través del sistema financiero debe integrar los canales operativos, financieros y de confianza identificados por la JERS (2020) y el BCE (2025a) con las características de amplificación específicas de la IA.

- *Etapa 1: El origen de la perturbación.* Un modelo de vanguardia —utilizado por un adversario, un actor estatal o incluso un usuario legítimo insuficientemente supervisado— identifica una vulnerabilidad en una dependencia de *software* crítica de una o más instituciones financieras. Calibración: Mythos encontró 271 vulnerabilidades en Firefox en una sola ronda de evaluación (The Next Web, 2026) e informa de miles de vulnerabilidades de alta gravedad en los principales sistemas operativos y navegadores (Anthropic, 2026a).

- *Etapa 2: La capa operativa.* La vulnerabilidad se convierte en un arma; los bancos sufren interrupciones en sus sistemas centrales, canales de pago o canales de atención al cliente. Calibración: El ataque de *ransomware* sufrido por el ICBC en noviembre de 2023 interrumpió liquidaciones del Tesoro por valor de más de 9.000 millones de dólares (Tesoro, 2024b); la interrupción de CrowdStrike del 19 de julio de 2024 afectó a aproximadamente 8,5 millones de dispositivos Windows (Microsoft, 2024), y solo Delta perdió 500 millones de dólares (Delta Air Lines SEC 8-K, 2024).

- *Etapa 3: El nivel financiero.* Los bancos incapaces de procesar pagos o cumplir con sus obligaciones comienzan a incumplir sus exposiciones interbancarias o a desencadenar ajustes de márgenes. Las contrapartes imponen condiciones más estrictas; comienza el acaparamiento de liquidez. Calibración: el saldo de caja negativo de SVB al cierre del 9 de marzo fue de aproximadamente 958 millones de dólares (CA DFPI, 2023); la liquidación de las operaciones de *carry trade* en yenes del 5 de agosto de 2024 supuso la liquidación de posiciones de *carry* a plazo por valor de aproximadamente 160. 000 millones de dólares a nivel mundial (Lombardi et al., 2024).

- *Etapa 4: La capa de confianza.* Las redes sociales y las comunicaciones amplificadas por la IA difunden información sobre la perturbación. Los depositantes y las contrapartes reaccionan. Calibración: 6.528 tuits sobre la «corrida» al SVB publicados entre el 8 y el 13 de marzo de 2023; los bancos en el tercil superior de exposición en Twitter perdieron 6,6 puntos porcentuales más de valor bursátil (Cookson et al., 2025).

- *Etapa 5: Transmisión macroprudencial.* Si los bancos experimentan pánico bancario correlacionado y tienen posiciones de activos impulsadas por IA correlacionadas (por ejemplo, coberturas similares basadas en modelos), la liquidación simultánea produce distorsiones en el mercado. Calibración: la caída repentina del 6 de mayo de 2010 vio cómo 4.100 millones de dólares en ventas algorítmicas provocaron una caída del 9 % en el DJIA en 36 minutos (CFTC y SEC de EE. UU., 2010).

La IA amplifica cada etapa: la característica de velocidad comprime las etapas 1 y 2; la característica de uniformidad correlaciona la etapa 3 entre instituciones; los retos de supervisión y la característica de velocidad amplifican la etapa 4; la característica de concentración convierte una perturbación por empresa en una perturbación sectorial.

7. ANÁLISIS DE ESCENARIOS: CUATRO SIMULACIONES CALIBRADAS

7.1. Escenario A: ciberataque impulsado por IA contra una G-SIB

- *Referencia de calibración*: El informe «IBM Cost of a Data Breach 2024» indica que el coste medio de una filtración en el sector de los servicios financieros es de 6,08 millones de dólares, con un coste medio de las megafiltraciones (más de 50 millones de registros) de 375 millones de dólares (IBM Security, 2024). Tiempo medio de detección y contención: 258 días. Sophos (2024) indica una tasa de ataques de *ransomware* del 65 % en los servicios financieros, con un rescate medio de 2 millones de dólares y un coste medio de recuperación de 2,58 millones de dólares. El FMI (GFSR de abril de 2024) proyectó unas pérdidas cibernéticas de 2.200 millones de dólares para una entidad financiera, con una frecuencia de una vez cada década.

- *Escenario adverso:* un actor de ataques autónomos de clase Mythos (ya sea patrocinado por un Estado o un grupo criminal con acceso de tipo «frontera» a través de filtraciones o extracción de modelos) lanza ataques multiobjetivo y multivectoriales contra los sistemas bancarios centrales de un G-SIB y tres de sus principales proveedores. Parámetros calibrados: 8,5 millones de terminales afectadas (análogo de CrowdStrike); interrupción del servicio de 8 a 24 horas (CrowdStrike se restableció en 78 minutos, pero la limpieza llevó días); ataque simultáneo a la infraestructura de liquidación del Tesoro (análogo del ICBC) que interrumpe transacciones por valor de más de 9.000 millones de dólares; demanda de rescate en el percentil 95 de los datos de servicios financieros de Sophos para 2024.

- I*mpacto previsto (ilustrativo, basado en puntos de calibración):* las pérdidas directas, incluyendo la interrupción del servicio, la reparación a los clientes y los costes de recuperación, se acercan a la estimación del FMI de 2.200 millones de dólares, una cifra que se da una vez cada década. Las pérdidas indirectas —multas regulatorias (multa de 80 millones de dólares de la OCC a Capital One; OCC, 2020), acuerdos de demanda colectiva (190 millones de dólares; Capital One Settlement Administration, 2022) y pérdida de negocio— podrían igualar o superar las pérdidas directas. El tiempo de recuperación supera

considerablemente el umbral tolerable de la prueba de resistencia cibernética del BCE para muchas instituciones (BCE, 2024b).

7.2. Escenario B: Retirada masiva de depósitos amplificada por la IA

- *Referencia de calibración:* SVB perdió 42.000 millones de dólares (aproximadamente el 25 % de los 166.000 millones de dólares de depósitos totales) el 9 de marzo de 2023, con otros 100.000 millones de dólares en solicitudes de retirada pendientes para el 10 de marzo (Junta de la Reserva Federal, 2023). A finales de 2022, los depósitos no asegurados ascendían a 151.600 millones de dólares, lo que representaba el 93,8 % del total (CA DFPI, 2023). Cookson et al. (2025) demuestran una caída excesiva de las acciones de 4,3 puntos porcentuales por cada desviación estándar de la exposición preexistente en Twitter, que se eleva a 6,6 puntos porcentuales para los bancos con exposición en el tercil superior.

- *Escenario adverso*: Un banco regional (aproximadamente 200.000 millones de dólares en activos, aproximadamente 60 % de depósitos no asegurados) sufre una campaña coordinada de desinformación generada por IA, que incluye un vídeo *deepfake* de pánico «filtrado» entre los ejecutivos internos, publicaciones masivas en redes sociales generadas por IA en varios idiomas y artículos de noticias sintéticos. La campaña se lanza fuera del horario laboral y se amplifica mediante redes coordinadas de bots en X, TikTok, WhatsApp y Telegram.

- *Impacto previsto. Calibrado en función de SVB, pero comprimido*: SVB sufrió una salida de depósitos del 25 % en una sola sesión bursátil; la proyección asume una salida del 25-40 % en 12 horas, teniendo en cuenta (i) la maduración de la infraestructura de banca digital desde 2023, (ii) el factor de amplificación del contenido generado por IA, y (iii) el precedente sentado por SVB. El Banco de Inglaterra (2025) señaló que el 70 % de los encuestados en la Encuesta de Riesgo Sistémico del Reino Unido del primer semestre de 2024 citó los ciberataques como un riesgo para el sistema financiero británico, lo que sugiere una conciencia generalizada de esta vulnerabilidad entre los supervisores.

7.3. Escenario C: Evento repentino de comportamiento gregario algorítmico

- *Referencia de calibración*: 6 de mayo de 2010, el DJIA cayó 998,5 puntos (aproximadamente un 9 %) en el transcurso de la jornada en 36 minutos; el desencadenante fue un programa algorítmico de venta de E-Mini por valor de 4.100 millones de dólares; el volumen total del 6 de mayo fue de 19.400 millones de acciones (CFTC y SEC de EE. UU., 2010). 5 de agosto de 2024: el TOPIX cayó un 12 % en un solo día (Boletín del BIS n.º 90; Lombardi et al., 2024), el Nikkei un 12,4 %, y el VIX superó brevemente los 60 puntos; alrededor de 160.000 millones de dólares en posiciones de *carry* a plazo de fondos de cobertura; recuperación en el plazo de una semana. El FMI (2024) informa de que el *trading* algorítmico representa aproximadamente el 70 % de la renta variable estadounidense y aproximadamente el 50 % de los futuros estadounidenses.

- *Escenario adverso:* una estrategia de negociación correlacionada impulsada por IA —por ejemplo, múltiples bancos y fondos de cobertura que utilizan superposiciones de sentimiento de noticias basadas en modelos de lenguaje grande similares en posiciones de un sector concentrado— recibe una noticia falsa generada por IA (quizás generada por un adversario). Los modelos desencadenan simultáneamente la venta.

- *Impacto previsto:* El escenario combina la velocidad de 2010 (algoritmos que reaccionan en milisegundos) con la escala de 2024 (posiciones transfronterizas similares al *carry trade*). Si los mercados responden de forma similar al episodio de 2024, son plausibles caídas bursátiles de un 10-12 % en un solo día en los índices afectados, con el VIX superando los 60 puntos. Los cortacircuitos, rediseñados tras 2010, limitan los peores resultados, pero el FMI (2024) recomendó explícitamente la recalibración de los cortacircuitos «a la luz de los movimientos de precios potencialmente rápidos impulsados por la IA».

7.4. Escenario D: Interrupción o efecto en cadena de un proveedor externo de IA

- *Referencia de calibración:* CrowdStrike (19 de julio de 2024) afectó a cerca de 8,5 millones de dispositivos Windows, con aproximadamente 24.000 clientes de CrowdStrike, incluyendo alrededor del 60 % de las empresas de la lista Fortune 500; solo Delta informó de unas pérdidas de cerca de 500 millones de dólares; Parametrix estimó unas pérdidas totales de 5.400 millones de dólares para las 500 principales empresas estadounidenses, excluyendo a Microsoft (Parametrix, 2024). Las interrupciones de los servicios en la nube a nivel mundial sumaron un total de 205,3 horas de interrupción en 2023, frente a las 133,5 horas de 2022 (Parametrix, el FMI, 2024).

- *Escenario adverso*: un importante proveedor de IA (Anthropic, OpenAI) o un hiperescalador (AWS, Azure, a CrowdStrike), un ataque dirigido (adversario de clase Mythos) o un fallo del modelo (envenenamiento de datos durante el entrenamiento). Dado que el 44 % de los modelos de IA de terceros del sector financiero del Reino Unido proceden de los tres principales proveedores (Breeden, 2024) y que AWS/Azure/GCP alojan conjuntamente aproximadamente el 63 % de las cargas de trabajo en la nube de las empresas (Synergy Research Group, 2025), la interrupción se propaga simultáneamente a múltiples G-SIB.

- *Impacto previsto:* los costes directos de la interrupción del servicio se escalan con estimaciones equivalentes a las de CrowdStrike. Fundamentalmente, la interrupción correlacionada en múltiples G-SIB convierte lo que sería un incidente operativo por empresa en uno que afecta a todo el sector. El régimen DORA CTPP es la principal medida de mitigación, pero las designaciones de Proveedores Externos Críticos aún no se han finalizado a mediados de 2026, e incluso tras la designación, el régimen se basa en recomendaciones y supervisión en lugar de en una autoridad prudencial directa.

TABLA 3. RESUMEN DEL ESCENARIO

Escenario	Fuente de calibración	Impacto principal (ilustrativo)	Medida de mitigación clave
A. Ciberataque contra G-SIB impulsado por IA	IBM 2024; Sophos 2024; ICBC 2023; FMI 2024	Pérdidas directas de entre 1.000 y 2.000 millones de dólares; interrupción de varios días	DORA TLPT; SR 26-2; Marco de estrés cibernético del BCE
B. Retirada masiva de depósitos amplificada por IA	SVB, marzo de 2023; Cookson et al., 2025	Salida de depósitos del 25-40 % en 12 horas	Informes de liquidez en tiempo real; reforma de la FDIC
C. Evento repentino de comportamiento gregario algorítmico	Crash del 6 de mayo de 2010; liquidación del 5 de agosto de 2024	Caída de la renta variable del -10 % al -12 %; VIX >60	Recalibración de los mecanismos de interrupción de la negociación; revisión de los márgenes
D. Cascada de proveedores de IA/hiperescaladores	CrowdStrike, julio de 2024; interrupción del servicio en la nube de 205,3 horas (2023)	Interrupción simultánea de múltiples G-SIB	Régimen DORA CTPP; simulacros de salida de la nube

8. ANÁLISIS EN PROFUNDIDAD ESPECÍFICO DE ESPAÑA

8.1. Postura cibernética y exposición a incidentes

INCIBE gestionó 122 223 incidentes de ciberseguridad en 2025, lo que supone un aumento interanual del 26 % (INCIBE, 2026). De los 401 incidentes que afectaron a operadores esenciales e importantes dentro del ámbito de aplicación de la NIS2, la banca representó el 34 %, la mayor cuota sectorial. El balance de 2024 registró 97.348 incidentes y 341 incidentes dentro del ámbito de aplicación de la NIS2.

El sector español de la ciberseguridad generó 6.351 millones de euros en ingresos en 2024 (+70 % desde 2020), con 164.761 profesionales de la ciberseguridad, lo que convierte a España en el cuarto mercado de ciberseguridad más grande de la UE (12 % de los ingresos continentales). Sin embargo, España no cumplió el plazo de transposición de la NIS2 del 17 de octubre de 2024, por lo que la Comisión Europea emitió un dictamen motivado en mayo de 2025 y la ley de aplicación aún no se había publicado a mediados de 2026.

8.2. Dependencias transfronterizas

Los bancos españoles están especialmente expuestos a la transmisión transfronteriza de IA/ciberseguridad. Santander cuenta con la designación de G-SIB (Banco de España, 2025a). Su composición de ingresos en 2024 muestra que Brasil representa el 21,6 %, España el 17,4 %, EE. UU. el 12,3 %, el Reino Unido el 11,2 %, México el 10,2 % y Polonia el 6,3 %. Santander Bank N.A. cuenta con 102.000 millones de dólares en activos en EE. UU. y 1,8 millones de clientes estadounidenses.

8.3. Capacidad de supervisión

El Banco de España cuenta con aproximadamente 3.475 profesionales, frente a los 6.968 de la Banca d'Italia, los 8.958 del Banque de France y los 10.255 del Bundesbank. El gobernador Escrivá ha señalado públicamente que el Banco de España dispone del 5 % de los recursos tecnológicos de sus homólogos más grandes del Eurosistema, a pesar de poseer el 12 % del capital del BCE. La AESIA, con cerca de 30 empleados en 2025, se enfrenta a un reto de escala similar en relación con sus homólogos de la UE y la amplitud de su mandato legal . La revisión por pares de España realizada por el Consejo de Estabilidad Financiera (18 de noviembre de 2025) recomendó desarrollar un panorama sectorial exhaustivo de las ciberamenazas, un análisis nacional de riesgos de terceros y una notificación de incidentes optimizada.

9. RECOMENDACIONES DE POLÍTICA

El caso Mythos ha puesto de manifiesto que la arquitectura regulatoria existente, aunque está bastante desarrollada, presenta lagunas críticas. A continuación se formulan cinco recomendaciones.

9.1. Para los supervisores microprudenciales

En primer lugar, revisar la exclusión de la IA generativa y la IA agentiva de la gestión del riesgo de modelos. La nota al pie 3 de la SR 26-2, que excluye la IA generativa y la IA agentiva de su ámbito de aplicación (Junta de la Reserva Federal et al., 2026), refleja el reconocimiento de que los principios tradicionales de la gestión del riesgo de modelos (MRM) no se adaptan bien a los modelos de lenguaje grande (LLM). Pero la alternativa no puede ser el silencio regulatorio. Los supervisores deberían publicar orientaciones provisionales —basándose en la norma NIST AI 600-1 (Autio et al., 2024), el seguimiento de la EBA sobre el aprendizaje automático en el IRB (EBA, 2023) y el dictamen de la EIOPA (EIOPA, 2025)— que establezcan unas expectativas mínimas para el uso de la IA generativa en actividades reguladas. La vía seguida por la UE en virtud de los artículos 9 a 15 de la Ley de IA ofrece un modelo útil.

En segundo lugar, armonizar las pruebas de resistencia cibernéticas entre jurisdicciones. La prueba de resistencia cibernética del BCE de 2024 (109 bancos, 28 en análisis en profundidad; BCE, 2024b) establece la norma operativa europea; los programas CBEST y SIMEX del Banco de Inglaterra (Banco de Inglaterra, 2022, 2023) constituyen el equivalente británico. Estados Unidos debería desarrollar un equivalente federal horizontal —basándose en el trabajo a nivel estatal del NYDFS— que se lleve a cabo anualmente.

9.2. Para las autoridades macroprudenciales

En tercer lugar, designar a los proveedores de IA críticos en el marco de regímenes similares a la DORA a nivel mundial. El régimen CTPP de la DORA (artículos 31-44; Reglamento Delegado (UE) 2024/1502 de la Comisión) es el marco más desarrollado. Las designaciones CTPP deben ultimarse con urgencia y ampliarse explícitamente para abarcar los laboratorios de IA de vanguardia, junto con los hiperescaladores de la nube. Estados Unidos debería establecer un régimen de designación paralelo, posiblemente a través de una acción del Consejo de Supervisión de la Estabilidad Financiera. El informe de buenas prácticas del FSB de 2026 debería proporcionar una plantilla internacional convergente.

En cuarto lugar, incorporar escenarios de pánico amplificados por la IA en los marcos de liquidez y resolución. La conclusión de Cookson et al. (2025) de que la exposición a las redes sociales durante el pánico del SVB provocó un exceso de pérdidas de entre 4,3 y 6,6 puntos porcentuales ha sido ahora validada empíricamente. Los coeficientes de liquidez calibrados según hipótesis de salidas de fondos anteriores a 2023 han quedado obsoletos. La FDIC, la Reserva Federal, el MUS del BCE y el Banco de España ( ) deberían actualizar conjuntamente las hipótesis de salida en el LCR y la planificación de la resolución para reflejar la dinámica de las retiradas masivas en la era digital.

9.3. Para el sector

En quinto lugar, se debe estandarizar la gestión de riesgos de los proveedores de IA y adoptar el marco del FS-ISAC como mínimo. Los seis libros blancos sobre IA del FS-ISAC (FS-ISAC, 2024) —en particular «Evaluación de proveedores de IA generativa» y «Evaluación cualitativa de riesgos»— proporcionan un estándar práctico desarrollado por el sector. Los bancos deberían adoptar estos (o equivalentes) como estándares mínimos, integrados con el RMF de IA del NIST (NIST, 2023) y los requisitos de evaluación de la conformidad de la Ley de IA de la UE.

9.4. Para los laboratorios de IA de Vanguardia

En sexto lugar, formalizar protocolos de divulgación responsable para las vulnerabilidades de la infraestructura financiera detectadas por la IA. El acuerdo a medida de Anthropic con el FSB —en el que se acordó informar, restringir la distribución más amplia a petición de la Casa Blanca (Financial Times, 2026) y mantener la lista de socios de Glasswing, compuesta por unas 40 organizaciones— improvisó un camino que debería institucionalizarse. Un protocolo debería especificar: (i) qué institutos nacionales (CAISI de EE. UU., AISI del Reino Unido) tienen acceso permanente a las evaluaciones previas al despliegue; (ii) qué organismos sectoriales (FS-ISAC, MSS del BCE, Banco de España) reciben información oportuna sobre los avances en las capacidades; (iii) cómo se amplía simétricamente el acceso a las capacidades defensivas (la asimetría de Mythos —40 organizaciones con acceso, miles sin él— es inestable); y (iv) cómo los plazos de divulgación equilibran el tiempo de parcheo y el tiempo de explotación.

Séptimo, ampliar la evaluación independiente previa a la implementación. Los benchmarks de duración de tareas de METR (Kwa et al., 2025), las evaluaciones de esquemas de Apollo Research (Apollo Research, 2024), las evaluaciones de capacidades cibernéticas del AISI del Reino Unido (UK AI Security Institute, 2026a) y los *benchmarks* de seguridad ofensiva de XBOW (XBOW, 2026) proporcionan, en conjunto, un ecosistema de evaluación en maduración. El RSP de Anthropic (Anthropic, 2025), el Preparedness Framework v2 de OpenAI (OpenAI, 2025) y el Frontier Safety Framework v3 de Google DeepMind (Google DeepMind, 2025) proporcionan

equivalentes desde el punto de vista del sector. La evaluación previa a la implementación debería ser obligatoria, en lugar de voluntaria, para los modelos de vanguardia ofrecidos a las instituciones financieras sistémicas.

10. CONCLUSIÓN

El asunto Mythos de mayo de 2026 es, a primera vista, una historia sobre un laboratorio de IA de vanguardia que colabora con un organismo normativo mundial en torno a las vulnerabilidades cibernéticas en la infraestructura de la estabilidad financiera. Más allá de la superficie, es una historia sobre tres cambios estructurales a los que la comunidad de estabilidad financiera debe enfrentarse ahora.

El primer cambio se refiere a la naturaleza misma del riesgo cibernético. Cuando un modelo de vanguardia puede detectar miles de vulnerabilidades de alta gravedad en cuestión de días y escribir *exploits* funcionales con una tasa de éxito del 83,1 % en el primer intento (Anthropic, 2026a), el tiempo necesario para explotar la vulnerabilidad se reduce por debajo del tiempo necesario para aplicar el parche, y los supuestos de equilibrio en los que se han basado décadas de políticas de resiliencia operativa quedan obsoletos. La proyección del FMI de que las pérdidas cibernéticas de una empresa financiera, que se producen una vez cada década, podrían alcanzar los 2200 millones de dólares (FMI, 2026) puede resultar en sí misma conservadora según los estándares de 2027.

El segundo cambio se produce en el perímetro regulatorio. Un puñado de laboratorios de IA de vanguardia e hiperescaladores en la nube —Anthropic, OpenAI, Google DeepMind, AWS, Azure, Google Cloud, NVIDIA— son ahora, desde el punto de vista operativo, tan fundamentales para la estabilidad financiera como los bancos más grandes, pero se encuentran fuera del perímetro prudencial. El régimen DORA CTPP ofrece el marco más desarrollado para cerrar esta brecha, pero se encuentra aún en una fase temprana de implementación. Estados Unidos no tiene ningún equivalente. El informe sobre buenas prácticas previsto por el FSB es la iniciativa política internacional más trascendental en este ámbito, y su diseño determinará las normas de la próxima década.

El tercer cambio se da en la política sobre la capacidad de la IA. El acuerdo de Anthropic de no distribuir Mythos más ampliamente a petición de la Casa Blanca (Financial Times, 2026) sienta un precedente: la capacidad de la IA de vanguardia es ahora una cuestión de política de seguridad nacional, además de política comercial. Como señaló el director ejecutivo de Mistral, Arthur Mensch, en mayo de 2026, «es imposible que Mythos compruebe el código fuente del ejército francés», y la implicación de una «dependencia irreversible» se extiende también a la infraestructura financiera (NL Times, 2026). La Comisión Europea, el Bundesbank y el Banco de España han pedido un acceso más amplio. Un régimen en el que los bancos españoles deban depender de la divulgación discrecional de un laboratorio estadounidense a un organismo internacional presidido por el Reino Unido es inestable.

En el caso concreto de España, las implicaciones políticas son graves. El Banco de España, la CNMV y la AESIA cuentan en conjunto con la mayor parte de las autoridades de supervisión que necesitan, pero con recursos relativamente limitados y una transposición de la NIS2 aún inconclusa. La condición de G-SIB de Santander y la profunda integración de OpenAI en el BBVA —que se extenderá a los 120 000 empleados en 25 países a finales de 2025 (OpenAI, 2025)— hacen que los bancos españoles estén inusualmente expuestos tanto a escenarios alcistas como bajistas. La elevada concentración del sistema bancario español (CR3 aproximadamente 72 %; Banco Mundial) significa que un incidente de la magnitud de Mythos que afectara a una o dos grandes instituciones podría convertirse, de forma plausible, en un evento sistémico.

La lección más profunda del momento Mythos es que la estabilidad financiera ha adquirido una nueva dependencia del comportamiento responsable de un pequeño número de actores no financieros. El marco tradicional de estabilidad financiera partía del supuesto de que las entidades de importancia sistémica relevantes eran los bancos (a los que se fueron sumando con el tiempo las cámaras de compensación de derivados, las ECC y algunas otras). El caso Mythos demuestra que los laboratorios de IA de vanguardia se encuentran ahora en este grupo, no porque acepten depósitos, sino porque sus decisiones sobre la capacidad de los modelos, su

lanzamiento y los protocolos de divulgación pueden afectar de manera significativa a la continuidad operativa de todos los bancos del mundo.

El Consejo de Estabilidad Financiera, la Reserva Federal, el Banco Central Europeo y el Banco de España necesitarán nuevas herramientas —o, más exactamente, nuevas aplicaciones de las herramientas existentes— para gestionar esta nueva dependencia. La primera generación de esas herramientas se está creando en estos momentos, en las salas de reuniones del FSB, en los borradores de consulta de los actos de ejecución de la DORA y en las tarjetas modelo de Claude Opus 5 y GPT-6. Que se desarrollen bien, con rapidez y de forma coordinada entre jurisdicciones es la cuestión central de la política de estabilidad financiera para lo que queda de década.

Referencias

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