LA INTELIGENCIA ARTIFICIAL EN EL SISTEMA FINANCIERO: IMPLICACIONES Y AVANCES BAJO LA PERSPECTIVA DE UN BANCO CENTRAL

Iván Balsategui, Sergio Gorjón y José Manuel Marqués

BANCO DE ESPAÑA

https://doi.org/10.53479/38235

Los autores pertenecen al Departamento de Innovación Financiera e Infraestructuras de Mercado, y agradecen los comentarios recibidos de José Manuel Carbó Martínez, Andrés Alonso-Robisco y un evaluador anónimo. Formulario de contacto para comentarios.

Resumen

La adopción del Reglamento de Inteligencia Artificial por parte de la Unión Europea, junto con la irrupción de los grandes modelos de lenguaje *[Large Language Models* (LLM)], basados en modelos fundacionales, o, de forma más genérica, de la inteligencia artificial generativa (GenAI, por sus siglas en inglés), ha despertado un renovado interés, dentro y fuera de la industria financiera, sobre las oportunidades y limitaciones que puede entrañar esta tecnología como elemento transformador de la sociedad. Este artículo tiene por objeto proporcionar el contexto en el que se han producido los avances recientes y ponerlos en perspectiva, identificar posibles hojas de ruta dentro del sistema financiero y enunciar, asimismo, los obstáculos y estímulos más importantes que podrían constreñir o favorecer su desarrollo a medio y a largo plazo.

Palabras clave: inteligencia artificial generativa, modelos fundacionales.

1 Introducción

La inteligencia artificial (IA) consiste, esencialmente, en el desarrollo de sistemas que realizan labores para las que típicamente se necesitan habilidades propias de los seres humanos. Su introducción cuenta con una larga tradición, si bien su evolución está ligada a avances recientes en la capacidad computacional y a cambios importantes en su enfoque. En concreto, estas técnicas han migrado de los modelos basados en reglas que se ejecutan de manera sistemática a otros de naturaleza probabilística. A diferencia de la econometría clásica, estos últimos no imponen una forma funcional específica sobre el origen de los datos. En su lugar, infieren directamente cuáles son los patrones subyacentes, mejorando así la capacidad de predicción o de toma de decisiones ante un determinado problema.

Esta evolución, unida a un abaratamiento en el coste del procesamiento y almacenamiento de la información y a una creciente digitalización de todo tipo de datos, es lo que ha hecho que crezcan el número de sectores y los casos de uso para los que la IA ofrece hoy soluciones. La industria financiera no ha sido ajena a este fenómeno y, en las últimas décadas, ha explorado la posibilidad de implementarlas en múltiples escenarios (por ejemplo, asesoramiento y análisis de riesgo de crédito) y procesos de gestión interna (por ejemplo, automatización) (Alonso-Robisco y Carbó, 2022).

Sin embargo, los desafíos para los usuarios en general, y para el sistema financiero en particular, son numerosos. En la práctica, estas soluciones suponen pasar de un escenario controlado, donde predominaba una implantación gradual y selectiva de los modelos basados en reglas, a otro en el que todas las áreas de negocio encuentran potenciales aplicaciones que desean llevar a producción cuanto antes, sobre todo tras irrumpir las propuestas de los

modelos de GenAI. Surge, así, la necesidad de contar con un marco que considere las implicaciones en términos de riesgo (incluyendo los posibles requerimientos regulatorios), analice y optimice la gestión de los recursos necesarios y, en última instancia, plantee las implicaciones estratégicas que conlleva su uso generalizado (Alonso-Robisco y Carbó, 2021).

Al aplicar estas técnicas, las autoridades y bancos centrales encaran los mismos dilemas que el sector privado, con la peculiaridad de que la comprensión de los riesgos y de sus mecanismos de contención adquiere una mayor relevancia dado su impacto potencial sobre la estabilidad financiera. En particular, bajo esta óptica se plantea una dicotomía entre, por un lado, las ventajas de favorecer una adopción acelerada de la IA y, por otro, cómo este hecho puede en sí mismo ser una fuente de desequilibrios para mercados como, por ejemplo, los del crédito. En concreto, por señalar tan solo algunos de los peligros más destacados, un uso profundo de la IA sin las debidas salvaguardas puede fácilmente dar lugar a riesgos de concentración cibernéticos, así como a los derivados de comportamientos de rebaño, con el consiguiente aumento de las correlaciones del mercado. Por tanto, a la hora de incorporar la IA para dar cumplimiento a sus respectivas obligaciones es preciso definir una hoja de ruta que sea consciente, mida estas implicaciones y priorice aquellas áreas con mayor impacto. Así, por ejemplo, hay campos como la prevención del blanqueo de capitales, la financiación del terrorismo o lo relativo a prevenir o detectar riesgos cibernéticos donde resulta más evidente su promoción a corto plazo.

Adicionalmente, a los bancos centrales les interesa conocer la repercusión que una adopción generalizada de la IA acabará teniendo en cuestiones como el mercado de trabajo, la productividad o el grado de desigualdad social. Los análisis preliminares sobre este impacto son hasta el momento muy heterogéneos. Algunos autores como, por ejemplo, Acemoglu (2024) consideran que el efecto es muy modesto, mientras que otros llegan a conclusiones contrarias (Cazzaniga, Pizzinelli, Rockal y Mendes, 2024). Se evidencia, así, un amplio rango de posibles resultados, como reflejan Aldasoro, Gambacorta, Korinek, Shreeti y Stein (2024), quienes proponen analizar sus efectos a partir de escenarios concretos, para concluir que la regulación deberá adaptarse según las transformaciones tecnológicas vayan materializándose en la economía real.

El presente artículo no aborda estos últimos aspectos, sino que se centra específicamente en el impacto directo de la IA en la industria financiera. Tras exponer los cambios que ha habido en el uso de los algoritmos, se identifican las actividades principales en las que sus efectos pueden llegar a ser significativos para, a continuación, hablar de los riesgos más importantes que entrañan y de las respuestas que están planteando tanto los reguladores como las autoridades sectoriales en general.

2 Un breve recorrido por el mundo de la inteligencia artificial[1]

A lo largo de su historia, la IA ha combinado períodos de intensa actividad y progreso con épocas de estancamiento e incluso abandono (inviernos). Ya en 1842 la matemática y pionera

1 Para un análisis más exhaustivo, véase Nayak y Walton (2024).

de la informática Ada Lovelace comenzó a vislumbrar que el recorrido de las computadoras iría más allá del simple procesamiento numérico (Carlucci Aiello, 2016). Posteriormente, Alan Turing publicó un artículo[2] donde se planteaba si las máquinas podrían pensar o imitar el comportamiento de la mente humana. Es aquí cuando se formula la famosa prueba para demostrar si una máquina es capaz de convencer a una persona de que está ante un ser humano[3]. No fue hasta 2014 cuando un programa informático, el bot conversacional Eugene Goostman, logró superar dicho test (Warwick y Shah, 2016).

Con independencia de lo anterior, el término «inteligencia artificial» surgió en la Conferencia de Dartmouth de 1956 para describir la ciencia y la ingeniería de hacer máquinas inteligentes[4] (Moor, 2006). De este modo, se logró formalizar el concepto y se abrió un nuevo campo de estudio científico. Comenzó así una primera edad de oro, que se extendió hasta 1974[5]. El contraste entre unas expectativas infladas sobre sus posibilidades y las limitaciones prácticas de la tecnología[6] originó su primer invierno, que dio paso a una nueva etapa de esplendor (1980-1987).

Este último período resultó particularmente fructífero e implicó, entre otras cosas, la llegada de los sistemas expertos que, basados en la programación explícita de reglas del formato «si… entonces…», eran capaces de capturar el conocimiento en determinadas materias para poder tomar decisiones informadas[7]. También supuso un nuevo paradigma en el entrenamiento de las redes neuronales para los modelos probabilísticos, donde estas fueron capaces de minimizar los errores a partir del aprendizaje de los datos de entrenamiento, impulsando el resurgimiento de la investigación del aprendizaje profundo *(deep learning).*

A finales del siglo XX y comienzos del XXI, se produjeron nuevos avances significativos en materia de IA impulsados, principalmente, por un incremento en la capacidad de computación, la aparición de algoritmos de aprendizaje automático *(machine learning)* más sofisticados y una mayor disponibilidad de datos para mejorar el entrenamiento de los modelos[8]. Por último, a partir de 2017 se asientan los fundamentos de una IA generativa, con la creación de una nueva arquitectura de redes neuronales *(transformers)* basada en un mecanismo de «atención» (Vaswani et al., 2017), que otorgaría a estos algoritmos un

2 "Computing Machinery and Intelligence" (1950).

3 Test de Turing.

4 Organizada por un grupo de científicos de la computación de la época, su génesis se atribuye a John McCarthy.

5 Algunos trabajos destacados de este período son el desarrollo por el psicólogo Frank Rosenblatt en 1957 del Perceptrón, una de las primeras redes neuronales capaces de reconocer patrones en imágenes y textos, y el desarrollo del primer bot conversacional, llamado ELIZA, desarrollado por Joseph Weizenbaum entre 1964 y 1966.

6 Por ejemplo, durante la Guerra Fría se confió en que la IA pudiera realizar traducciones automáticas, lo que no fue posible. Del mismo modo, los resultados de los experimentos con redes neuronales no fueron los esperados.

7 Tuvieron mucha popularidad en esa época y, como se verá más adelante, suponen el núcleo de lo que se llegará a denominar la IA simbólica.

8 Algunos ejemplos ilustrativos son la creación del programa de ajedrez Deepblue de IBM, que venció al campeón de ajedrez Gary Kasparov en 1997; la presentación del sistema Watson por IBM, con capacidad de responder a preguntas en lenguaje natural y que venció a los concursantes del popular concurso americano Jeopardy!, o el lanzamiento en 2016 de AlphaGo, que con el uso de redes neuronales profundas y técnicas de aprendizaje por refuerzo venció al campeón mundial de go, Lee Seedol, en una histórica partida a cinco juegos.

renovado potencial. Para ello se parte de unos modelos entrenados con un amplio conjunto de datos heterogéneos que, como resultado, pasan a ser capaces de realizar una variedad de tareas generales (modelos fundacionales) como, por ejemplo, el caso de GPT *(Generative Pre-trained Transformer),* que puede comprender el lenguaje natural. Posteriormente, sobre estos modelos se entrenan los algoritmos que generan de forma autónoma texto, imágenes o vídeos como respuesta a las preguntas que el usuario formula a través de un *prompt*[9].

La transición descrita supone, en esencia, la evolución de la IA desde enfoques analíticos hacia modelos generativos o, lo que es lo mismo, de aplicar reglas lógicas y programación más específica para replicar la inteligencia humana (la llamada IA simbólica) a tratar de emular el funcionamiento del cerebro humano a nivel más estructural (IA conexionista) (véase esquema 1). Un claro ejemplo de esto es el empleo, dentro del sistema financiero, de modelos de IA simbólica como los usados por los departamentos de *marketing* para determinar aspectos como la propensión de compra de ciertos activos financieros por parte de los clientes de las entidades según su historial, o anticipar la pérdida de un cliente de una entidad de forma que se puedan realizar labores preventivas *(churning).* Por otro lado, el sistema financiero también está utilizando sistemas de IA conexionista para disponer de modelos que permitan detectar el fraude de una forma más óptima dentro de las áreas de cumplimiento y fraude, para desarrollar bots conversacionales como herramienta de soporte interno para la gestión de la cartera de clientes por parte de los empleados del departamento comercial, así como para contar con modelos generativos que aumenten la eficiencia en la ingeniería de *software* a la hora de escribir y documentar código o escribir casos de prueba, entre otros usos.

3 El interés del sistema financiero en la inteligencia artificial: una instantánea

En los últimos años (sobre todo, tras la irrupción de la pandemia y la consiguiente transición hacia un nuevo escenario digital) la IA se ha ido consolidando como un soporte sustancial de la cadena de valor de la industria financiera[10]. Este desarrollo presenta diferencias a nivel geográfico y entre instituciones, si bien siempre ha estado ligado a la propia naturaleza del negocio bancario, muy centrado en la explotación y el análisis de la información.

9 Uno de los hitos que, sin duda, ha contribuido a este nuevo *hype* de la IA generativa se remonta a 2022, cuando la compañía OpenAI lanzó al público general ChatGPT, una aplicación de chatbot de inteligencia artificial (muy superior a la ELIZA de 1966) basada en el módulo fundacional GPT, que se ha convertido en la aplicación de Internet con el crecimiento más rápido de la historia. En apenas dos meses llegó a tener más de cien millones de usuarios activos, un logro que otro tipo de tecnologías como TikTok tardaron nueve meses en alcanzar (o hasta dos años y medio, en el caso de Instagram).

10 Sin ir más lejos, para el año 2022 la Autoridad Bancaria Europea cifraba en casi un 80% el número de bancos de la región que declaraban hacer un uso proactivo de herramientas de IA con diferentes propósitos. Si se tenía en cuenta, además, al colectivo de entidades que estaban teniendo debates o que contaban con desarrollos incipientes en este campo, el porcentaje se elevaba a más del 98% (EBA, 2023). No obstante, otras fuentes basadas en indicadores cuantitativos parecen apuntar a unas tasas de adopción efectivas aún bajas.

Esquema 1

Tipología de IA

En función del grado de desarrollo de los modelos, de la generalidad o no de los algoritmos que se desarrollan y de la capacidad que tenga la IA de superar o no el razonamiento humano en todas las facetas de la vida, la IA se puede clasificar en tres tipologías diferentes:

01

IA estrecha o débil

El diseño y el funcionamiento de los modelos están pensados para **realizar una sola tarea** de la mejor manera posible.

Es una forma de IA diseñada específicamente para centrarse en una sola tarea en concreto. Por ejemplo, en sistemas de reconocimiento facial y de imágenes, chatbots y asistentes conversacionales (Google Assistant, Siri, Alexa), vehículos autónomos, modelos predictivos, motores de recomendación, simuladores de juegos como el ajedrez, go, etc.

02

IA general o fuerte

Todavía está en desarrollo… El sistema de IA podría igualar e incluso superar a la inteligencia humana en **múltiples tareas cognitivas.**

Este tipo de IA sería capaz de adaptarse y comprender nuevos contextos sin programación específica, es decir, el sistema aprendería de sí mismo. Para ello el propio sistema debería embeber habilidades del tipo razonamiento, aprendizaje automático, comprensión del lenguaje natural y la resolución de problemas en multiples dominios.

Es el tipo de IA en el que se está trabajando actualmente para el futuro.

03

Superinteligencia artificial

Cuando la IA supera ampliamente a la inteligencia humana y podría **evolucionar** y tomar decisiones de manera **autónoma y** sin su pervisión.

El hecho de que los modelos de IA alcancen una IA general o fuerte podría conllevar la multiplicación de esa inteligencia de forma exponencial a través de su propio aprendizaje autónomo. Esto se conseguiría gracias a un proceso denominado de **superación personal recursiva** *(recursive self-improvement)*, que haría que la IA mejorase continuamente en unos tiempos inalcanzables para los humanos.

Quizá sería ese el momento que los científicos denominan **singularidad,** es decir, cuando la IA supere a la inteligencia humana.

FUENTE: Comisión Europea (2024).

En el caso del sistema financiero, además de los factores tecnológicos descritos en el apartado anterior, han tenido un papel importante elementos idiosincrásicos como, por ejemplo, la contracción sostenida de los márgenes empresariales y, por extensión, la necesidad de obtener nuevas fuentes de eficiencias e ingresos o la competencia ejercida por las *fintechs* (Boukherouaa et al., 2021).

Del mismo modo, la aparición de los modelos fundacionales de GenAI ha dado un impulso renovado a la adopción de herramientas que facilitan un acceso asequible a modelos preentrenados de propósito más amplio[11], y habilitan, además, fórmulas de interacción usuario-máquina basadas en el lenguaje natural más que en el dominio de nociones de programación.

A semejanza de lo ocurrido en otros sectores, las razones por las que la IA ha despertado el interés de las entidades financieras son múltiples y diversas, y están asociadas principalmente a la promesa de alcanzar mejoras de productividad, reducciones de costes operativos e incrementos en la calidad y seguridad de productos, servicios y procesos. En la misma línea, en el apetito de estos jugadores ha tenido mucho que ver su potencial para optimizar la rentabilidad de las inversiones, aumentar las tasas de satisfacción de los clientes o profundizar en los niveles de inclusión financiera (Fernández, 2019). En este último sentido, la IA no solo puede ayudar a complementar una evaluación de crédito allí donde otros métodos tradicionales presentan limitaciones, sino que también puede capitalizarse para aumentar los niveles de bancarización, en la medida en que preste asistencia en la cumplimentación de los trámites formales o en la selección de los proveedores de servicio más adecuados para cada caso, aspectos que, con frecuencia y junto con el coste, son los factores que constituyen los obstáculos principales para el acceso.

En este contexto, y sin ánimo de ser exhaustivos, el uso de herramientas de aprendizaje automático por parte de la banca es especialmente notable en algunos frentes para, por ejemplo:

— Facilitar el cumplimiento de las obligaciones regulatorias a las que está sujeta (por ejemplo, reportes estadísticos, prevención del blanqueo de capitales o la financiación del terrorismo), favoreciendo así una menor incidencia de errores gracias a la mecanización del flujo de trabajo, la observancia de los plazos de entrega y la remisión de información de mayor calidad, claridad, granularidad y precisión.

— Optimizar, de forma más general, los procesos de negocio internos y tratar así de materializar ahorros de coste, bien mediante la reducción del número de tareas manuales que se realizan y liberando recursos para otras actividades, bien incrementando el rendimiento de la fuerza de trabajo con la prestación de asistencia informada en los procesos que gestiona (por ejemplo, el asesoramiento a los clientes).

— Contribuir a ejercer un control más eficaz y diligente de los riesgos bancarios, tanto en lo que respecta a la dimensión operativa (por ejemplo, detección y reacción temprana ante posibles fraudes o equivocaciones) y financiera (por ejemplo, evaluación y seguimiento de la solvencia y probabilidad de impago o predicción de

11 Lo que supone un ahorro sustancial en términos de la inversión necesaria para desarrollar y ajustar internamente los modelos, los costes asociados a su entrenamiento y el *time-to-market* estimado para completar con éxito su despliegue efectivo.

los flujos de caja)[12] como a los aspectos relacionados con la seguridad cibernética (por ejemplo, reducir las tasas de falsos positivos e identificar anomalías o correlaciones no triviales).

— Incrementar la capacidad analítica y predictiva en relación con eventos del mercado al objeto de maximizar el rendimiento de las inversiones en entornos de volatilidad incierta. Estas herramientas posibilitan la realización de estimaciones más precisas y robustas, que se nutren de datos no estructurados, procedentes de fuentes no tradicionales e incorporan información en tiempo real para generar un conocimiento que permita la toma de decisiones de forma dinámica[13].

— Mejorar la experiencia de los usuarios, lo que incluye la promoción y personalización de los productos y servicios que se comercializan para alcanzar mayores niveles de satisfacción y de retención de clientes, pero también la agilización de los procesos de pre y posventa, así como la ampliación de la accesibilidad general a los servicios bancarios a través de la automatización de ciertos canales de interacción (chatbots).

Pese a la amplitud aparente de esta casuística, el grueso de las iniciativas que en la actualidad están en producción se concentra en torno a aplicaciones de *back* y *middle office*. Se trata de una decisión lógica guiada por la intención de contener eventuales riesgos asociados a deficiencias o controversias en el uso de esta tecnología con clientes finales (Aldasoro, Gambacorta, Korinek, Shreeti y Stein, 2024), en un entorno en el que todavía no se ha consolidado e implementado completamente el desarrollo normativo en el que se enmarque el uso de estas técnicas.

Por su parte, los bancos centrales y las autoridades sectoriales también han mostrado un claro deseo de adoptar herramientas de IA con miras a mejorar el ejercicio de sus funciones[14]. A lo anterior se une, además, su interés por dilucidar, a través de la experimentación en primera persona, las oportunidades y riesgos que les son inherentes y, así, estar en mejor situación para comprender y evaluar su impacto real sobre la industria financiera.

Así las cosas, el acceso continuado a grandes cantidades de información de naturaleza compleja y granular, unido al incremento de los recursos tecnológicos a su disposición

12 Las ganancias potenciales incluyen la reducción de posibles pérdidas (Khandani, Kim y Lo, 2010), la optimización del capital (Fraisse y Laporte, 2022), la automatización de la toma de decisiones (Owolabi, Uche, Adeniken, Ihejirika, Bin Islam y Chhetri, 2024) o la expansión de la cifra de negocio a través de una ampliación de la base de solicitantes de crédito cuyas peticiones son aprobadas (Sadok, Sakka y Maknouzi, 2022).

13 A modo ilustrativo cabe señalar cómo las GenAI tienen la capacidad para reforzar los mecanismos de descubrimiento de precios o reducir las barreras de entrada a mercados de activos menos líquidos, como los de deuda corporativa o los emergentes.

14 Tanto los roles clásicos a los que están asociados —supervisión micro y macroprudencial, supervisión de conducta, vigilancia de los sistemas de pago o análisis económico— como algunos más recientes que están apareciendo: innovación e inclusión financiera o protección medioambiental (Carstens, 2019).

(Cipollone, 2024) —así como la satisfacción de otras metas estratégicas[15]—, han sido las grandes claves de esta transformación. De ahí que sean, precisamente, áreas tales como la estadística, el análisis macroeconómico, la vigilancia de los sistemas de pago o la supervisión donde se han producido los avances más destacados (Araujo, Doerr, Gambacorta y Tissot, 2024).

Comenzando por esta última, los beneficios de la IA se materializan en tanto que, en un contexto marcado por riesgos cambiantes y proveedores de servicios bancarios más heterogéneos, estas herramientas sirvan verdaderamente para potenciar la mayor efectividad de unas actuaciones orientadas a controlar la solidez, la solvencia y el comportamiento de las entidades financieras en consonancia con el objetivo de salvaguardar la estabilidad de todo el sistema (Beerman, Prenio y Zamil, 2021). Así, la detección temprana de riesgos latentes mediante la captura de anomalías/señales ocultas en los datos reportados o el incremento de la precisión a la hora de simular e identificar las consecuencias de escenarios adversos *(stress testing),* son ventajas latentes que todas estas autoridades coinciden en investigar.

Del mismo modo, esa mayor capacidad analítica redunda a favor de las actividades de modelización dirigidas a establecer la naturaleza, escala e incidencia potencial de desequilibrios de carácter estructural en la economía gracias, en particular, a su efectividad a la hora de identificar comportamientos no lineales (Hellwig, 2021). En concreto, la IA ayuda a establecer, con mayor acierto, los patrones que pueden subyacer en las relaciones entre variables sin llegar, no obstante, a inferir su causalidad[16] y, por extensión, a construir modelos más sofisticados y representativos del comportamiento de los agentes (Atashbar y Shi, 2023). La explotación de datos granulares, no estructurados[17] y accesibles de forma continua en el tiempo resulta fundamental para facilitar la construcción de indicadores económicos relevantes que permitan el despliegue de políticas públicas cuando estas resulten más efectivas (Doerr, Gambacorta y Serena-Garralda, 2021).

En la misma línea, la IA abona el terreno para profundizar en la monitorización de los circuitos de pago. En consecuencia, apoya la búsqueda de patrones inesperados en las transacciones que se intercambian con la finalidad de poder alertar de la presencia de actividades fraudulentas o ilícitas, o con el objeto de anticipar problemas de liquidez u operativos ya sea a nivel de entidades individuales o en el conjunto del sistema (Rubio, Barucca, Gage, Arroyo and Morales-Resendiz, 2020). Aunque su implantación es por el momento limitada, estos modelos facilitan la identificación de canales por los que se pueden transmitir perturbaciones

15 Por ejemplo, en el caso del Banco de España, el Plan Estratégico 2020-2024 contemplaba expresamente como una de sus prioridades la de impulsar la innovación tecnológica dentro de la institución. Se quería así promover su modernización, para lo que se hizo especial hincapié en favorecer la transformación digital, la gestión integrada de la información y la administración del riesgo de ciberseguridad. El desarrollo y despliegue de herramientas *suptech* fue solo una de las acciones que se desplegaron a estos efectos.

16 Especialmente en los casos de no linealidad, como evidencia Bahrammirzaee (2010).

17 Como, por ejemplo, el uso de técnicas de *web scraping* para recolectar información sobre precios, la explotación de datos de satélites como *proxy* de la evolución de la actividad económica o el análisis de microdatos de carácter no económico que empresas y particulares comparten a diario a través de las redes sociales (Shabsigh y Boukherouaa, 2023).

de naturaleza sistémica, lo que ayuda a contrarrestar sus potenciales efectos al acelerar las medidas de respuesta y dirigirlas a la auténtica raíz del problema.

Potenciar la calidad y utilidad funcional del dato es otro de los campos en los que el valor diferencial de la IA es más evidente. Este objetivo ha ganado importancia a medida que aumentaba el volumen, el detalle, la complejidad y la frecuencia de la información, promoviendo así la implantación de procesos de validación automática para detectar errores, observaciones atípicas u omisiones relevantes (Araujo, Bruno, Marcucci, Schmidt y Tissot, 2022). Al margen de utilizarse para identificar deficiencias en la información, estas herramientas pueden emplearse también para tratar de corregir los efectos negativos vinculados con la propia recopilación de la información, como es el caso de las deficiencias asociadas a muestras de tamaño reducido[18]. Mediante la combinación de distintas técnicas, el aprendizaje automático posibilita la limpieza, imputación y modelado de los datos faltantes, incluyendo su interpolación. Esto permite reforzar la robustez de los modelos y, eventualmente, prevenir el sobreajuste derivado que se puede producir en la fase de entrenamiento (Rebuffi, Gowal, Calian, Stimberg, Wiles y Mann, 2021).

Finalmente, la IA ofrece un amplio recorrido para hacer más efectivas las interacciones de los bancos centrales con el público. Esto incluye, por ejemplo, la adaptación de su estrategia de comunicación, cuidando el lenguaje y contenido utilizados (Bholat, Broughton, Parker, Ter Meer y Walczak, 2018), la ampliación de su alcance fruto de las oportunidades de traducción automática que ofrecen algunos de sus más recientes desarrollos (Cipollone, 2024) y la simplificación de la normativa de cuya redacción sea responsable (Moreno, Gorjón y Hernáez, 2021).

En lo que respecta al despliegue de soluciones basadas en modelos de GenAI el escenario que se dibuja es distinto, pues su alcance, posibilidades e implementación, como hemos comentado, requieren una aproximación estratégica y un mayor grado de coordinación, al tiempo que resulta necesario gestionar los retos y riesgos que se detallan en el siguiente epígrafe. En su mayor parte, los desarrollos actuales se limitan a pruebas o pilotos, quedando su aplicación confinada a casos de uso internos, orientados a la mejora de procesos de negocio. En general, estas herramientas se ejecutan en entornos aislados del exterior por razones de seguridad y protección de datos sensibles, ciñéndose su aplicación a la condición de asistentes virtuales con los que buscar potenciar la capacidad de toma de decisiones por parte de los seres humanos (OCDE, 2019).

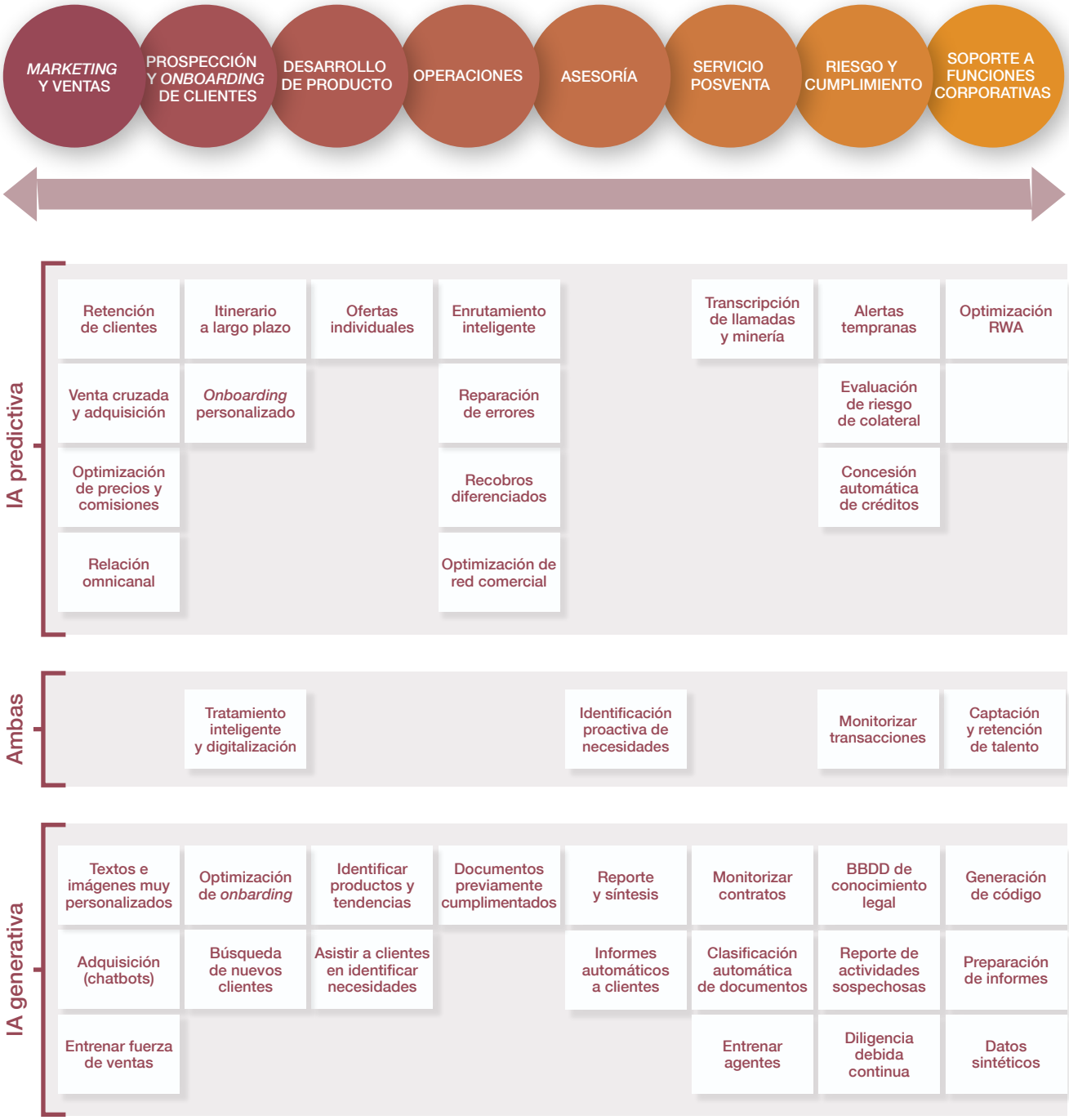
Entre otros ejemplos, destaca lo relacionado con la generación automática de códigos de programación, así como la realización de pruebas y depuración de los ya existentes. Del mismo modo, la elaboración de resúmenes de texto o de traducciones automáticas cuentan, a día de hoy, con un nutrido número de iniciativas.

Finalmente, son cada vez más las entidades que capitalizan estas herramientas para transcribir y analizar en tiempo real las conversaciones con clientes y asistir a los gestores en la resolución

18 Los sesgos o el ruido, entre otros.

Esquema 2

Áreas de aplicación de la IA en la industria financiera



FUENTE: Elaboración propia a partir de Riemer et al. (2023).

de incidencias o en la comercialización de los productos y servicios mejor adaptados. No obstante, el recorrido potencial de la GenAI va mucho más allá y, una vez solventados aspectos claves relacionados con su gobernanza y con la gestión de los datos, es previsible que su huella aumente y se extienda a otros ámbitos con rapidez (véase esquema 2).

En este sentido, ya se han documentado iniciativas dirigidas a optimizar el esfuerzo de *marketing* y ventas tanto a nivel de la comunicación como del diseño de la oferta comercial. Se persigue, así, acomodar ambos aspectos a la realidad individual de cada cliente. Del mismo modo, la mejora de la capacidad para identificar posibles fraudes o actividades ilícitas se está combinando con la cumplimentación automática de los informes asociados. Dicho esto, donde más valor diferencial puede aportar esta variante en un futuro posiblemente sea en las actividades de *front office,* toda vez que termine por hacer sus interacciones indistinguibles de las humanas.

4 Riesgos y barreras más destacados

Como ya se mencionó en el epígrafe 2, 2022 supone el punto de inflexión para el resurgimiento de la GenAI. Durante este período se han publicado diversos LLM, como ChatGPT3, ChatGPT4, Claude, Gemini, Llama o Mistral, entre otros, que ponen de manifiesto sus avanzadas capacidades de procesamiento del lenguaje natural (NLP, por sus siglas en inglés), de generación de textos e imágenes y, en definitiva, de contenido creativo.

El impacto transformador que puede tener esta tecnología moderna se puso pronto de manifiesto a pesar del estadio inicial en el que se encuentra. En este sentido, la carrera de las compañías por ofrecer una ventaja competitiva con este tipo de soluciones debería ir acompañada de una gestión de su riesgo intrínseco. En Estados Unidos, ya en 2023 el National Institute of Standards and Technology publicó su Marco de Gestión de Riesgos de la IA (AI RMF 1.0) para ayudar a las empresas a gestionar de forma eficiente dichos riesgos.

En un artículo de Gartner publicado en 2023 sobre los requerimientos para usar la IA de forma segura y efectiva se destaca: «En 2026, los modelos de IA de las organizaciones que instrumentalicen la transparencia, la confianza y la seguridad de la IA lograrán una mejora del 50 % en términos de adopción, objetivos de empresa y aceptación del usuario».

Sin perjuicio de que los modelos más clásicos de IA presenten también retos en esta dirección (Alonso-Robisco y Carbó, 2022), y dada la novedad y recorrido potencial de la IA generativa, en esta sección nos centraremos en los riesgos específicos de su empleo. En concreto, esta reviste diferentes tipos de problemas potenciales, entre los que destacan:

— *Exactitud en la salida de los modelos:* según el artículo publicado por McKinsey en mayo de 2024, titulado “The state of AI in early 2024: Gen AI adoption spikes and starts to generate value”, el principal riesgo que las organizaciones encuentran a la hora de explotar los modelos de GenAI es la falta de precisión de los mismos. ¿Qué ocurre cuando un sistema de GenAI genera texto, imágenes y otros resultados que son inexactos, incorrectos, engañosos o inapropiados? Este resultado puede deberse a un problema de calidad de los datos con los que se ha entrenado al modelo, o incluso venir motivado por los ajustes del propio sistema o por la falta de explicabilidad inherente de algunos modelos de GenAI. Todo ello puede conllevar

que la toma de decisiones sea errónea, con las implicaciones que esto pudiera tener para el proceso de negocio de cada organización[19]. Para mitigar estas situaciones resulta de especial relevancia disponer de entornos o marcos de pruebas que permitan comprobar el correcto funcionamiento de los dos componentes más relevantes en un sistema de GenAI: el que proporciona la información de contexto *(retriever)* y el que conforma y genera la respuesta en sí *(generator).*

— *Alucinaciones de los modelos:* este fenómeno se produce cuando un modelo genera información que aparentemente parece plausible, pero que en realidad es inventada y no se basa en la información de los datos de entrenamiento del modelo. Debido a la propia naturaleza de los modelos LLM, estos siempre devolverán una secuencia de texto salvo que se indique expresamente lo contrario. Para evitar estas alucinaciones existen tres estrategias complementarias: técnicas de ingeniería de construcción del *prompt* con el que se consulta el modelo, uso de técnicas como las *Retrieval Augmented Generation* (RAG) y/o realizar un proceso de calibración o *fine-tuning* del propio modelo, con sus pros y contras (véase esquema 3).

— *Privacidad de la información:* A este respecto, se deberían abarcar las siguientes situaciones:

  • El principal motor de la GenAI son los datos que, en un mundo cada vez más digitalizado, existen de modo creciente. El problema radica en que es posible que parte de la información que se emplee para entrenar a los modelos pudiera contener información personalmente identificable (PII, por sus siglas en inglés). En consecuencia, el simple hecho de utilizar estos datos para el entrenamiento o que estén incorporados al resultado podría estar exponiendo detalles confidenciales sobre las personas y conllevar la correspondiente violación de la privacidad y posibles usos indebidos. A día de hoy existen en el mercado diferentes soluciones y técnicas para mitigar este riesgo[20].

  • Asimismo, existe una técnica de ataque sobre los modelos LLM denominada *Membership Inference Attack* que permite conocer los datos originales empleados para entrenar los modelos de IA (incluyendo información personal) sin tener información ni de la arquitectura del modelo ni de los parámetros que se han usado dentro del propio modelo, lo que podría suponer una violación de la privacidad de la información (Shokri et al., 2017).

---

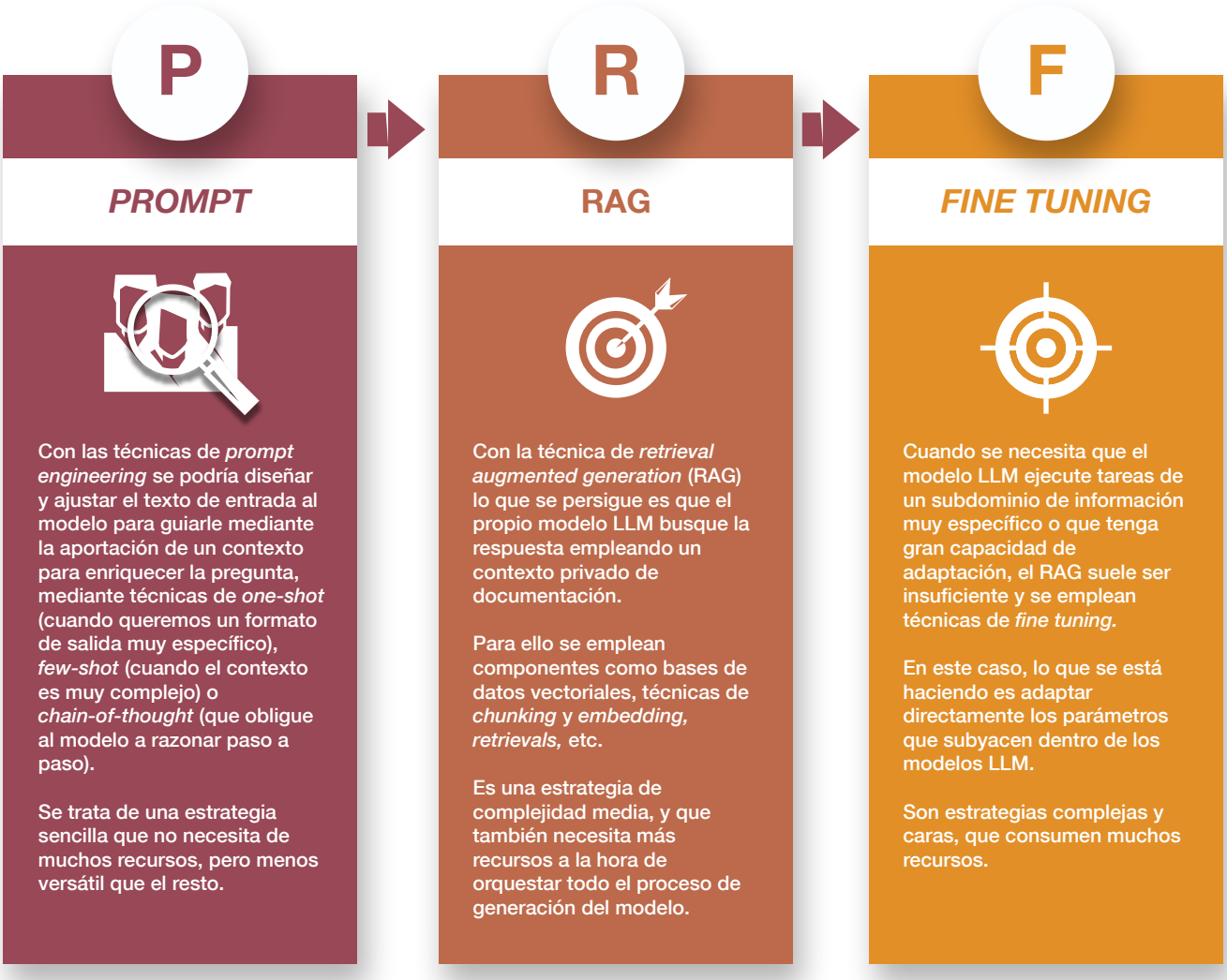
19 A modo de ejemplo, merece la pena señalar el caso del magnate hongkonés Li Kin-kan, que en 2017 llegó a perder hasta 20 millones de dólares al día a través de un bot de inversión llamado K1, por las decisiones que el propio modelo de GenAI generaba en función de diversas fuentes de información con las que trabajaba y que posteriormente trasladaba a operaciones de mercado. El magnate demandó a la empresa de inversión Tyndaris Investment, ubicada en Londres, que comercializaba el uso de este bot de inversión inteligente.

20 Google, por ejemplo, dispone de una solución llamada *data loss prevention* para garantizar que toda la PII se identifique y anonimice automáticamente antes de entrenar los modelos. Asimismo, pueden emplearse técnicas como *differential privacy, homomorphic encryption, secure multi-party computation* o *federated learning* para preservar la privacidad a la hora de emplear modelos de IA. Otro concepto que también se está usando a la hora de preservar la privacidad es el de *data privacy vault,* como el servicio Protecto Vault de la empresa Protecto AI, que usa tokenización.

Esquema 3

Estrategias para reducir las alucinaciones de los modelos LLM

Con objeto de reducir el riesgo de alucinación en el resultado que los modelos LLM pueden generar se plantean tres posibles estrategias a seguir, desde la más sencilla, como sería el *prompt engineering,* pasando por la aplicación de un *retrieval augmented generation* (RAG), hasta la más compleja a seguir, como podría ser hacer un *fine tuning* del propio modelo en sí.



FUENTE: Elaboración propia.

- Asimismo, cuando por error, negligencia o de forma intencionada un usuario de estos modelos de GenAI pública escribe o pega información confidencial en un *prompt,* automáticamente esta información pasaría a estar disponible al creador/gestor de estos modelos y, potencialmente, para el resto de los usuarios[21].

21 En 2023 sucedió un caso de un ingeniero de Samsung que subió código sensible de la empresa a ChatGPT, lo que supuso automáticamente una revelación de información confidencial a terceros, con el consiguiente impacto reputacional en la compañía.

— *Derechos de autor, propiedad intelectual:* como se mencionó anteriormente, los modelos de GenAI emplean ingentes cantidades de datos para su entrenamiento. Podría darse el caso de que dentro de esta cantidad de información existiera material protegido por derechos de autor, lo que conllevaría cumplir con la normativa vigente en esa materia a la hora de poder usar dicha información. Si el propietario del modelo de GenAI no cumple con esta normativa, estaría infringiendo los derechos de propiedad intelectual. También la generación de nuevo contenido que pudiera ser muy parecido a obras ya existentes podría abrir el camino a disputas legales sobre su originalidad y propiedad ante la existencia de posibles plagios[22].

— *Resultados sesgados:* la mayoría de los modelos de GenAI se entrenan, principalmente, con datos de calidad más que cuestionable extraídos de Internet. Es importante resaltar que un modelo de GenAI puede generar resultados sesgados, en general originados a partir de unos datos de entrenamiento que también contienen dicho sesgo. Las compañías deberían priorizar el uso de fuentes de datos confiables y de alta calidad a la hora de entrenar sus modelos de GenAI como primer paso para mitigar este problema, evitando en la medida de lo posible conjuntos de datos que pudieran incluir enfoques de tipo racista[23] o sexista, entre otros. Además, la opacidad de los modelos de GenAI no ayuda, precisamente, a mitigar este riesgo. No obstante, han surgido corrientes como la IA explicable (XAI, por sus siglas en inglés) que abordan los temas de interpretabilidad y explicabilidad de un modelo (Arrieta et al., 2019).

— *Seguridad de los modelos:* la seguridad aplicada a este tipo de modelos de GenAI abarca un gran abanico de posibilidades. Si bien se puede pensar en la utilización de estos modelos para reforzar, en primera instancia, las labores de protección de los sistemas actuales de las organizaciones, también presenta otra disyuntiva que es la que emplean los *hackers* para sofisticar técnicas de ataque a las infraestructuras de terceros haciendo uso de estos modelos. A modo de resumen, a continuación se enumeran los riesgos de seguridad más relevantes:

  • El nivel de sofisticación que mediante el empleo de la GenAI se puede dar en los ataques de *phishing* a usuarios se ha elevado exponencialmente dada la calidad de la redacción (tono, estilo, formato) y contenido de los textos de los correos electrónicos maliciosos enviados a terceros. Los problemas de mala gramática o redacción son suplidos con creces por modelos de GenAI, haciendo que la diferenciación entre un correo legítimo y otro que no lo es haya disminuido sustancialmente. Asimismo, los atacantes, mediante el empleo de la GenAI,

22 Un claro ejemplo de ello ha sido la demanda que el *New York Daily News,* el *Denver Post* y el *Chicago Tribune* han interpuesto contra las compañías OpenAI y Microsoft por infracción de derechos de autor.

23 Una muestra es lo que le ocurrió en 2015 al modelo desarrollado por Amazon de reclutamiento de desarrolladores de *software,* donde se partía de unos datos de entrenamiento sesgados, según los cuales se asignaba un mayor peso a los currículums procedentes de hombres que de mujeres, lo que sesgaba el resultado final del modelo.

pueden escribir, desarrollar, perfeccionar y depurar *malware* capaz de eludir las medidas de seguridad tradicionales.

- Dado que, en general, los sistemas de GenAI se integran con variedad de diversas fuentes de datos, *application programming interfaces* y otros sistemas, y que la integración suele resultar compleja, la superficie de ataque puede resultar más extensa y generar vulnerabilidades que los atacantes podrían aprovechar para acceder a información confidencial de la compañía. Resultan especialmente relevantes las recomendaciones de la organización OWASP para identificar las principales vulnerabilidades de uso de LLM, que las organizaciones pueden tener en cuenta para minimizar este riesgo de ataque (véase esquema 4).

- Otro punto interesante son los *deepfakes*. Sin duda alguna los modelos actuales de GenAI son capaces de generar imágenes manipuladas, vídeos, voz sintética suplantada y toda clase de contenido hiperrealista que puede resultar muy complicado diferenciar de los contenidos reales, lo que podría provocar una difusión errónea de información con fines intencionados e incluso querer manipular a la opinión pública[24]. Estos temas son un debate ético que se debería de contemplar en todo proceso de gobernanza de riesgos de modelos de GenAI.

- No hay que olvidar que, independientemente de los posibles ataques externos que pueden sufrir este tipo de modelos, también hay que cuidar los sistemas de seguridad internos para evitar que un empleado pueda utilizar técnicas de exfiltración para «sacar» información fuera de la organización en beneficio propio (por ejemplo, empleando la técnica de *prompt injection*).

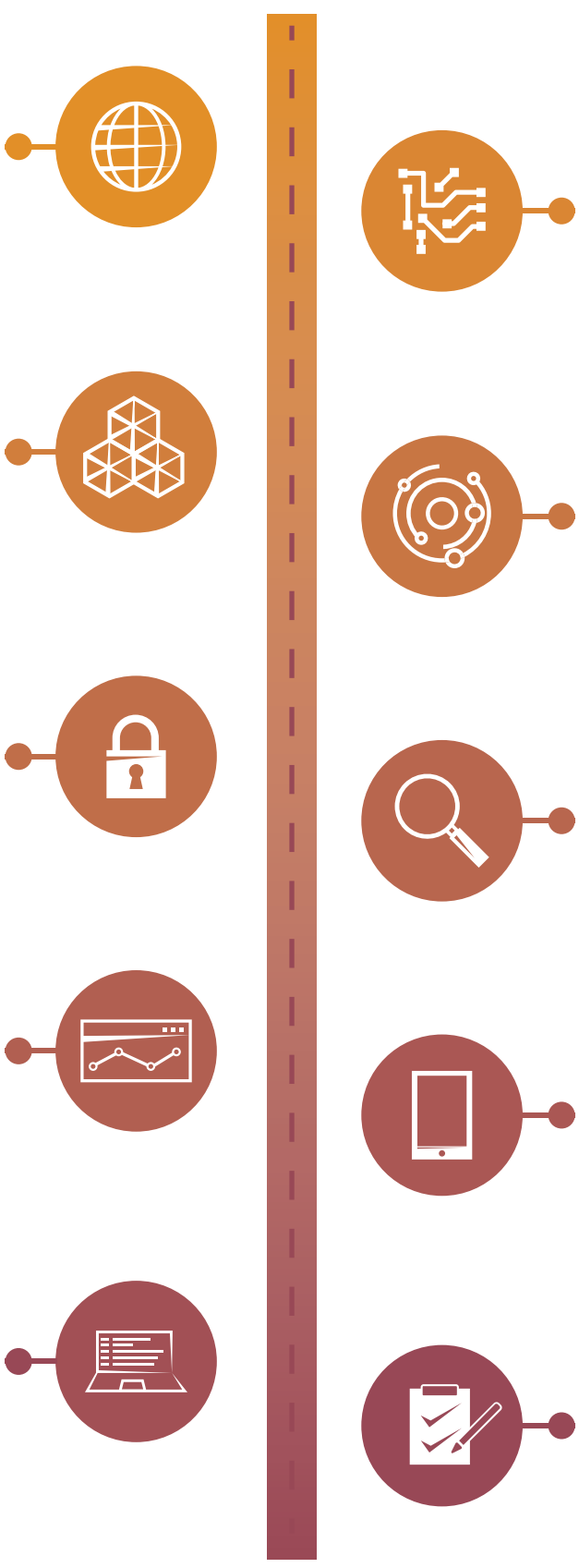
- Otro punto que se debería considerar es evaluar la manera en que los modelos de LLM han sido entrenados y con qué información, y evitar así que estos dispongan de «puertas traseras» que pudieran explotarse para permitir el acceso discrecional de un atacante, de forma ilegal, a los sistemas donde se encuentran alojados los modelos de LLM.

— *Conformidad con terceros:* resulta fundamental que cuando se diseñe un modelo de GenAI que maneje sobre todo información confidencial, y este se encuentre desplegado en la infraestructura de un proveedor de servicios de IA externo, dicho proveedor deberá disponer de certificaciones de cumplimiento para, al menos, asegurarnos de que *a priori* no se esté vulnerando ninguna norma de cumplimiento y protección de datos y que los posibles usos de los modelos de GenAI no sirvan para reentrenar a estos.

24 Un ejemplo podría ser, en 2024, la suplantación de la voz de Joe Biden, presidente de Estados Unidos, en una llamada telefónica realizada para desincentivar el voto en las primarias de Nuevo Hampshire.

Esquema 4

Principales vulnerabilidades del uso de LLM



FUENTE: OWASP (2023).

— *Coste de la experiencia y la computación:* al desarrollar, explotar y mantener modelos de GenAI se debería de tener en cuenta que estos sistemas requieren de *hardware* especializado, generalmente servicios en la nube, lo que puede suponer un gasto considerable. Además, dado que se trata de una tecnología «relativamente nueva», los perfiles tales como científicos de datos, ingenieros de aprendizaje automático y *prompt engineering*, entre otros, demandan salarios superiores. Si a esto se suma que la escasez de dichos perfiles es significativa, la situación presenta importantes barreras de entrada para muchas organizaciones.

No cabe duda de que toda tecnología emergente y disruptiva, como es el caso de la GenAI, conlleva riesgos inherentes en su desarrollo y uso, de tal manera que las compañías han de identificar, evaluar y mitigar anticipadamente, para poder estar alineadas con todos y cada uno de los aspectos anteriormente mencionados.

Por otro lado, se deberían tener en consideración cuáles son los principales retos técnicos a la hora de implementar de una forma efectiva la IA en las organizaciones.

Previo al resurgimiento de la GenAI, cuando las organizaciones se enfrentaban a proyectos de aprendizaje automático *(machine learning),* se seguía un paradigma llamado MLOps para mejorar los procesos en los que se desarrollan, implementan y monitorizan las aplicaciones basadas en aprendizaje automático. En este entorno de trabajo, científicos de datos, ingenieros de modelos de *machine learning,* ingenieros de datos, expertos en desarrollo e integración continua, así como en seguridad y privacidad, trabajaban conjuntamente desde la definición de los procesos de calidad de datos al entrenamiento de los modelos, el despliegue en repositorios centrales y la puesta en explotación de los sistemas.

Sin embargo, con la llegada de la GenAI este paradigma se extiende al propio concepto de MLOps, acuñándose un nuevo término, llamado LLMOps, centrado principalmente en la gestión, implementación y mantenimiento de los LLM. Debido a su complejidad y los requerimientos de recursos, los LLM plantean desafíos únicos en términos de operaciones, tales como la selección de modelos fundacionales, tareas de *prompting*, *benchmarking* de resultados de los modelos bajo un nuevo prisma de indicadores ajenos a los empleados en *machine learning*, procesos de *fine-tuning*, gobernanza de modelos, observabilidad, entre otros.

Sin duda, todos estos nuevos paradigmas que las organizaciones tienen que implementar no hacen más que engrosar la lista de complejas tareas a la hora de integrar los modelos LLM en sus sistemas propietarios. A esto se le unen nuevas necesidades computacionales, no solo a la hora de entrenar los modelos, sino también a efectos de su puesta en producción, bien sea vía *on-premises*, bien a través de proveedores en la nube, lo que sin duda va a resultar clave a la hora de tener éxito en su despliegue tanto interno como externo a la organización.

Cuadro 1

Aspectos destacados de las principales iniciativas regulatorias regionales

Europa	Estados Unidos	Asia
Reglamento europeo de carácter transversal de aplicación tanto a los proveedores de sistemas de inteligencia artificial susceptibles de ser utilizados dentro de la UE (con independencia de su origen geográfico) como a sus respectivos usuarios Establece requisitos (o, en su caso, prohibiciones) sobre los sistemas de IA que son proporcionales al riesgo que se deriva del uso que se les vaya a dar (no se regulan tecnologías concretas) Las obligaciones a las que se someten los proveedores incluyen, entre otras, cuestiones relacionadas con sistemas de gestión de calidad, elaboración de documentación técnica, registro de los sistemas o evaluaciones de conformidad Las obligaciones a las que se someten los usuarios incluyen, entre otras, cuestiones relacionadas con la adecuada supervisión humana, información sobre incidentes graves o malfuncionamiento y cumplimiento de otras exigencias legales como las relativas a la *General Data Protection Regulation* (GDPR)	En 2020 se crea una oficina de ámbito nacional responsable de la vigilancia e implementación de la estrategia del país en este campo Aún no existe normativa federal que regule, prohíba o restrinja el desarrollo y uso de la IA En su lugar se han promulgado diferentes tipos de actos jurídicos en distintos ámbitos que abordan facetas puntuales de su aplicación La Casa Blanca ha promovido algunas medidas, con especial atención a aspectos como el acceso y uso equitativo de sistemas de IA o el desarrollo de modelos fundacionales y cuestiones anejas de seguridad Sigue abierto el debate en torno a iniciativas legislativas que tocan, por ejemplo, los sistemas de decisión automática (transparencia, derecho de no participación, no discriminación), recogen la posibilidad de exigir licencias a determinados prestadores de servicios o protegen la propiedad intelectual, entre otros	Singapur cuenta con un marco de gobierno de la IA que promueve la explicabilidad, transparencia, equidad y salvaguarda de los derechos civiles Ha publicado, además, unas orientaciones relativas a la explotación de datos personales por la IA También está ultimando un conjunto de recomendaciones en materia de GenAI para fomentar un ecosistema de confianza La Autoridad Monetaria de Singapur (MAS, por sus siglas en inglés) está desarrollando un marco de gestión de riesgos de la GenAI en el sistema financiero China ha aprobado normas, orientadas a tecnologías concretas, que abordan diferentes tipos de riesgos de la IA Hay disposiciones que regulan el uso de algoritmos de recomendaciones, prohibiéndolos con menores y dando opción al usuario a no ser objeto de los mismos Otras disposiciones regulan a los proveedores y usuarios de tecnologías capaces de generar contenidos sintéticos (etiquetándolos y restringiendo ciertas aplicaciones) Recientemente se han promulgado normas en materia de GenAI que preservan derechos de propiedad intelectual o exigen la aplicación de medidas para garantizar la calidad, precisión y fiabilidad de los datos

FUENTE: Elaboración propia.

5 Respuestas de política pública y regulatoria: algunas consideraciones

Teniendo en cuenta los riesgos mencionados antes, y el impacto que puede tener la IA en la sociedad, no es de extrañar que las autoridades hayan comenzado a desarrollar un marco normativo y de política pública dirigido a su mitigación. Muchas de las actuaciones, tanto a escala nacional como internacional, tienen un carácter general, afectando por extensión al sistema financiero, pero sin que, por el momento, sean muchas las regulaciones específicas orientadas a este ramo de actividad. Aunque existen diferencias tanto en el detalle como en el nivel de exigencia, todas estas iniciativas reflejan objetivos comunes y, en última instancia, aspiran a conseguir que el despliegue de esta tecnología pueda producirse de la manera más ordenada posible (véase cuadro 1).

Uno de los primeros focos de atención es el relacionado con la dimensión ética de la IA o, dicho de otra manera, con proporcionar un marco que permita construir ecosistemas responsables para asegurar que los resultados de la IA sean justos, inclusivos, sostenibles y no discriminatorios (UNESCO, 2021). A estos efectos, uno de los mayores referentes internacionales son las Recomendaciones de la OCDE de 2019[25] que, a su vez, dieron pie a los Principios del G20 en la materia. Su objetivo ha sido ofrecer un estándar global específico, de carácter complementario a otros más genéricos ya existentes e igualmente aplicables a este terreno[26], que pueda servir de guía a las correspondientes actuaciones de las autoridades jurisdiccionales.

Las Recomendaciones de la OCDE se desarrollan a través de un conjunto de postulados que persiguen facilitar la ejecución de las políticas nacionales, además de fomentar la cooperación transfronteriza[27]. En lo que respecta a los Principios, estos animan a invertir en la investigación y el desarrollo de una ciencia abierta que propicie el intercambio de conocimiento e información. También defienden la creación de un marco de gobernanza y unas políticas públicas que estimulen la innovación y favorezcan la transición hacia escenarios de puesta en producción. Además, se espera que los gobiernos presten especial atención al reciclaje de la fuerza laboral y que busquen fomentar el diálogo interdisciplinar, dentro y fuera de sus fronteras, para favorecer grandes acuerdos.

La universalidad de estas premisas ha hecho que las mismas tengan su eco en numerosas iniciativas como, por ejemplo, en las Directrices Éticas de la UE para una IA confiable o, posteriormente, en el Reglamento de la UE de Inteligencia Artificial. Algo parecido puede decirse de otras jurisdicciones relevantes en este terreno como Estados Unidos[28], China[29], Japón[30] o Reino Unido[31], por citar algunas. Pese a la existencia de aproximaciones divergentes entre todas ellas en lo que respecta al detalle, hay coincidencia en reconocer la importancia de mantener un diálogo abierto que favorezca la cooperación y apueste por ajustar las exigencias prácticas al nivel de riesgo efectivo correspondiente a cada aplicación concreta de la IA.

En este escenario, Singapur marca, asimismo, una diferencia clara al significarse como uno de los países pioneros en reinterpretar dichas recomendaciones para facilitar su aplicación a

---

25 Desde entonces, han tenido lugar varias revisiones de las Recomendaciones con el fin de facilitar su implementación y poder acomodar los cambios técnicos y políticos que se han ido produciendo como, por ejemplo, los derivados de la irrupción de la IA generativa y, así, mantener su vigencia.

26 En materia de privacidad y protección de datos, seguridad de la información digital y conducta, principalmente.

27 Entre otros aspectos, abarcan dimensiones como las relativas a que la IA contribuya a la mejora del bienestar social, a que su aplicación sea transparente para los usuarios o que se garantice la robustez de los modelos basados en estas herramientas.

28 *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (2023). Por otro lado, comenzando por la *Algorithmic Accountability Act* de 2022, parece que se está perfilando una hoja de ruta definitiva en el Congreso estadounidense que podría sentar las bases para la puesta en marcha de un esfuerzo conjunto que permita ir dando forma a un paquete de iniciativas regulatorias sobre aspectos puntuales de la IA.

29 Caracterizada por un despliegue progresivo de normas centradas en aspectos concretos del uso de la IA que, en cada ámbito, tocan facetas relacionadas con la ética. Destacan las disposiciones que regulan los algoritmos de recomendaciones (2021), las que se ocupan de los contenidos generados sintéticamente —en prevención del fenómeno de los *deepfakes*— (2023) o la ley sobre IA generativa (2023). Este conjunto de iniciativas regulatorias parte, entre otros, del marco general de gobernanza que, en 2019, sentó las bases de los principios aplicables a una nueva generación de IA responsable (Sheehan, 2023).

30 *Social Principles of Human-Centric AI* (2019).

31 *Ethics, Transparency and Accountability Framework for Automated Decision-Making* (2023).

la industria financiera, donde cuenta ya con unos principios específicos[32]. Precisamente, esta parece ser la ruta preferida de los operadores del sector que reclaman unos criterios concretos que les proporcionen la necesaria certidumbre sobre la validez de las implementaciones que han de realizar.

Otra de las piezas definitorias del marco emergente de políticas públicas en torno a la IA es el ya señalado Reglamento de la UE. Con esta iniciativa, Europa se adelanta al resto de países en cuanto a la regulación de los potenciales usos de la IA con el propósito de reforzar su autonomía estratégica en el desarrollo del mercado digital y, especialmente, de salvaguardar el bienestar de la sociedad. Para ello adopta un enfoque consistente con los valores que caracterizan a la región, poniendo así especial acento en la defensa de los derechos humanos (Calderaro y Blumfelde, 2022).

Esta norma tiene carácter transversal y afecta tanto a los proveedores de sistemas de IA susceptibles de ser utilizados dentro de la UE (con independencia de su origen geográfico) como a sus respectivos usuarios, es decir, a quienes los exploten. El concepto de IA que maneja es bastante amplio para, así, dar cabida a avances como los de los modelos fundacionales[33]. Asimismo, el Reglamento establece una jerarquía de riesgos en función del destino de estas herramientas, lo que permite identificar una serie de categorías a las que imponer obligaciones proporcionadas (véase esquema 5)[34].

En lo que respecta al sistema financiero, resultan de particular importancia los sistemas calificados de alto riesgo y, en concreto, los relacionados con la identificación biométrica[35] y la calificación crediticia[36]. Estos, como cualquier otro que esté encuadrado en dicha categoría, habrán de someterse a una evaluación de conformidad y serán objeto de ciertas exigencias particulares.

Entre otros aspectos relevantes, los mencionados sistemas deberán garantizar: i) el despliegue de políticas de gestión de riesgo específicas; ii) la implementación de un modelo de gobernanza y gestión de datos de entrenamiento y prueba que sea robusto; iii) la preparación de documentación técnica que arroje evidencias suficientes acerca del cumplimiento de los requisitos; iv) el fomento de la transparencia y trazabilidad de las decisiones, y v) su adecuada supervisión a través de asegurar la intervención humana. Por otro lado, los sistemas de riesgo específico para la transparencia deberán poner en conocimiento del usuario el hecho de que la interacción se está produciendo con una máquina y no con una persona.

---

32 *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector.*

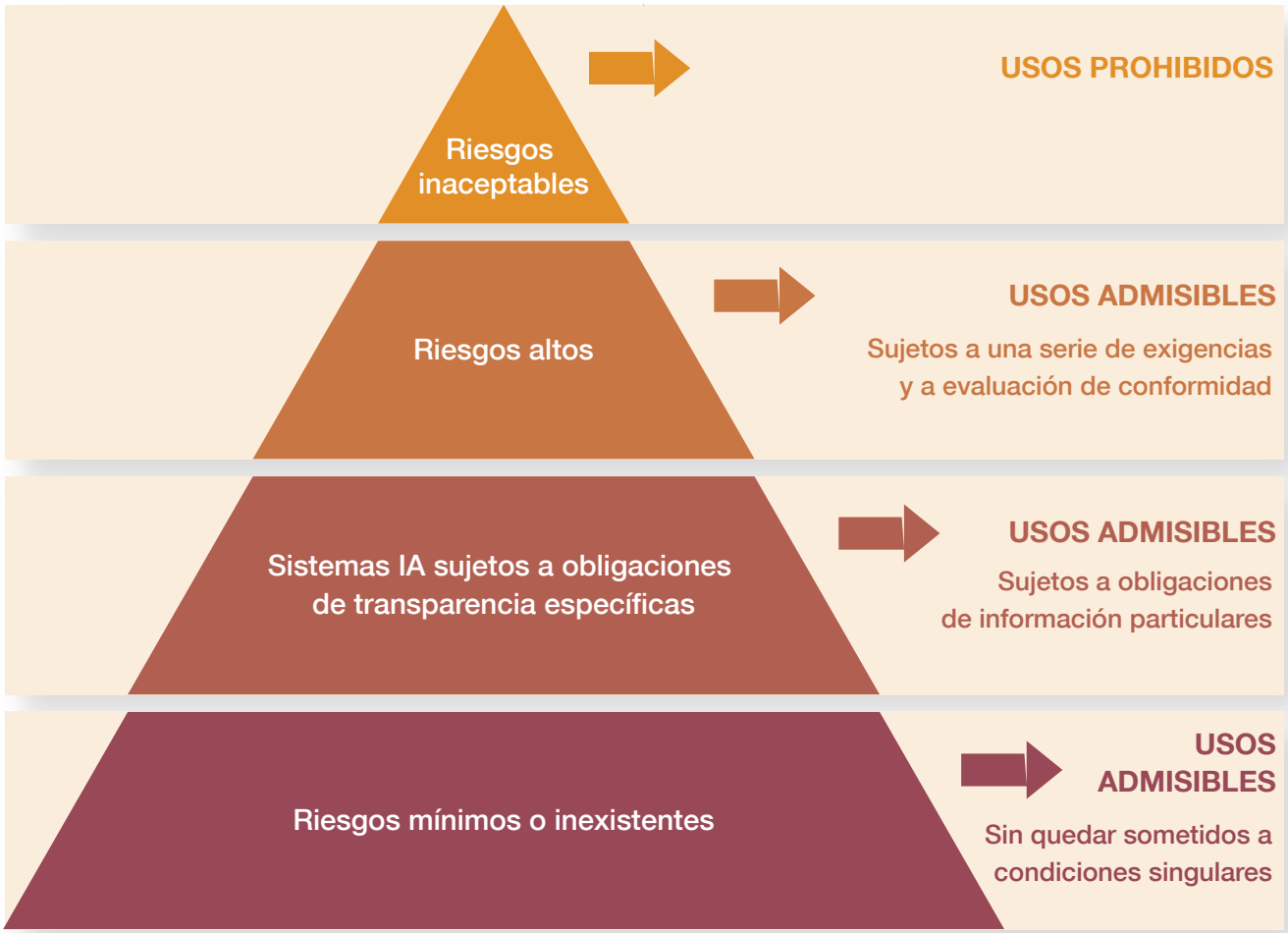
33 Durante la fase de negociación de la norma esta circunstancia ha sido objeto de cierto debate ante la posibilidad de que una definición tan abierta pudiera incluir, igualmente, los modelos de inferencia más tradicionales. No obstante, finalmente parece existir un consenso sobre que no es el caso.

34 Los potenciales usos abusivos de la IA como, por ejemplo, los que puedan distorsionar el comportamiento de las personas o los que conduzcan a la discriminación en el trato se consideran prohibidos.

35 Sin perjuicio de lo anterior, esta seguirá quedando sujeta al Reglamento General de Protección de Datos [Reglamento (UE) 2016/679], así como a la Directiva de tratamiento de datos personales por las autoridades [Directiva (UE) 2016/680].

36 La evaluación de riesgos y la fijación de precios en el sector de los seguros son otras de las áreas afectadas.

Esquema 5
Tipología de riesgos de la IA en el Reglamento de la UE



FUENTE: Comisión Europea (2024).

Si bien los principios en los que se sustenta el reglamento delimitan el terreno de juego general, corresponde a la futura normativa secundaria la tarea de clarificar su vertiente más práctica. Esto, unido a cómo de efectivos resulten los mecanismos que se han previsto para coordinar a las diferentes autoridades supervisoras, serán sin duda piezas clave para asegurar que el reglamento se implementa con la máxima coherencia, que evita fricciones innecesarias con la legislación sectorial existente y que exhibe la flexibilidad suficiente para permitir el desarrollo de una tecnología que, en gran medida, aún puede considerarse incipiente.

A este mismo objetivo se espera que contribuya el piloto de *sandbox* regulatorio promovido por el Ministerio para la Transformación Digital y de la Función Pública y que busca proponer orientaciones y buenas prácticas para facilitar la mejor aplicación de la norma. Aunque el reglamento es de carácter transversal, tal y como se ha comentado, existen varios casos de uso que afectan significativamente al sistema financiero. En este sentido cabe destacar que, recientemente, la Comisión Europea ha lanzado una consulta pública con objeto de dar encaje

de la manera más coherente a las implicaciones de esta regulación en el sistema financiero, dadas las conexiones de esta normativa con otras como las relativas a externalización o control de riesgos operativos.

Por último, mencionar que la divergencia existente en los enfoques seguidos por cada una de las distintas jurisdicciones internacionales es vista como un factor de fragmentación potencial en una industria que, por su propia naturaleza, no conoce fronteras. Por esta razón, crecen las iniciativas que, como el reciente acuerdo entre la OCDE y el Global Partnership on Artificial Intelligence, buscan avanzar en una agenda de acciones convergentes.

6 Conclusiones

Dejando a un lado a las empresas tecnológicas, la industria financiera es quizá el sector que, con mayor amplitud y profundidad, está sabiendo aprovechar las enormes oportunidades que ofrece la IA. Como parte de un itinerario dirigido a transformar la información en conocimiento y este en inteligencia, las entidades transitan con rapidez de técnicas más clásicas, basadas en enfoques analíticos, a otras que apuestan por replicar el comportamiento humano a nivel más estructural. En consecuencia, a día de hoy, el tipo de problemas que permiten abordar estos vehículos se antoja cada vez más complejo.

En este sentido, actualmente la IA es uno de los factores determinantes para que las entidades puedan alcanzar sus objetivos más habituales, tales como el aumento de la productividad, la mejora de su eficiencia, la reducción de los costes o el incremento de la calidad o seguridad de sus productos y servicios. En particular, en este último ámbito, el despliegue de estas herramientas se antoja urgente como vía para combatir el fraude o las amenazas cibernéticas, toda vez que crece el número de delincuentes que, valiéndose de estas mismas técnicas, están consiguiendo ya réditos importantes. Adicionalmente, una nueva generación de herramientas basadas en GenAI se está abriendo paso de modo progresivo con un impacto y alcance significativos y con potencial de modificar no solo los procesos internos, sino también la forma en que las organizaciones se relacionan con clientes y empleados.

Este escenario abre la puerta al desarrollo y explotación de fuentes complementarias de ingresos y habilita fórmulas novedosas que ayuden a maximizar la rentabilidad operativa en un entorno marcado por los cambios. No obstante, el calado real de esta contribución resulta aún poco visible, dada la cautela con la que las entidades están llevando a cabo a día de hoy su despliegue, centrándolo, en gran medida, en los procesos y aplicaciones internos. Es de esperar que conforme se asienten e implementen los avances legislativos el impacto sea mucho más disruptivo.

De modo similar, las autoridades financieras pueden también beneficiarse del recorrido potencial de estas tecnologías en lo que se refiere al desempeño de sus competencias y, por ende, contribuir positivamente a la mejora del bienestar social. De ahí que muchas de ellas

cuenten ya con programas y estrategias ambiciosos para promover su exploración y, eventualmente, facilitar su implantación.

No obstante, la generalización en el uso de estas técnicas implica, igualmente, la aparición de riesgos importantes cuya gestión no puede obviarse. En este sentido, el principal desafío práctico al que se enfrentan autoridades y usuarios consiste en desplegar un modelo de gobernanza apropiado y robusto que asegure la transparencia y seguridad de la tecnología. Solo así se conseguirá que gocen de la suficiente confianza entre la población como para facilitar su adopción y aceptación masivas.

De ahí que proliferen las iniciativas orientadas a establecer un marco regulatorio y supervisor que favorezca un despliegue lo más ordenado posible. Se trata, en resumidas cuentas, de conseguir que, se aplique donde se aplique, la IA ofrezca siempre resultados justos, inclusivos, sostenibles y no discriminatorios; algo de lo que el sistema financiero es particularmente tributario. En este sentido, quizá uno de los aspectos en los que es necesario avanzar más en el futuro es en la consecución de cierto grado de convergencia de los esfuerzos normativos a nivel internacional, con objeto de evitar que un fenómeno que por su propia naturaleza tiene una dimensión global se vea lastrado por una multiplicidad de normas fragmentarias de base nacional.

BIBLIOGRAFÍA

Acemoglu, Daron. (2024). "The simple macroeconomics of AI". *Economic Policy.* https://doi.org/10.1093/epolic/eiae042

Aldasoro, Iñaki, Leonardo Gambacorta, Anton Korinek, Vatsala Shreeti y Merlin Stein. (2024). "Intelligent financial system: how AI is transforming finance". BIS Working Papers, 1194, Bank for International Settlements. https://www.bis.org/publ/work1194.pdf

Alonso-Robisco, Andrés, and José Manuel Carbó. (2021). "Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation". Documentos Ocasionales, 2105, Banco de Espala.

Alonso-Robisco, Andrés, and José Manuel Carbó. (2022). "Inteligencia artificial y finanzas: una alianza estratégica", Documentos Ocasionales, 2222, Banco de España.

Araujo, Douglas, Sebastian Doerr, Leonardo Gambacorta y Bruno Tissot. (2024). "Artificial intelligence in central banking: Executive Summary". BIS Bulletin, 84, Bank for International Settlements. https://www.bis.org/publ/bisbull84.pdf

Araujo, Douglas, Giuseppe Bruno, Juri Marcucci, Rafael Schmidt y Bruno Tissot. (2022). "Machine learning applications in central banking". Irvin Fisher Committee on Central Bank Statistics Bulletin, 58, Bank for International Settlements. https://www.bis.org/ifc/publ/ifcb57_01_rh.pdf

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila y Francisco Herrera. (2019). "Explainable artificial intelligence (XAI): Concepts". *Taxonomies, Opportunities and challenges toward responsible AI*, 11. https://doi.org/10.1016/j.inffus.2019.12.012

Atashbar, Tohid, y Rui Aruhan Shi. (2023). "AI and Macroeconomic Modeling: Deep Reinforcement Learning in an RBC Model". IMF Working Paper, 40, International Monetary Fund. https://doi.org/10.5089/9798400235252.001

Bahrammirzaee, Arash. (2010). "A comparative survey of artificial intelligence applications in finance. Artificial neural networks, expert system and hybrid intelligent systems". *Neural Computing & Applications*, 19, pp. 1165-1195. https://doi.org/10.1007/s00521-010-0362-z

Beerman, Kenton, Jermy Prenio y Raihan Zamil. (2021). "Suptech tools for prudential supervision and their use during the pandemic". Financial Stability Institute Insights on Policy Implementation, 37, Bank for International Settlements. https://www.bis.org/fsi/publ/insights37.pdf

Bholat, David, Nida Broughton, Alice Parker, Janna Ter Meer y Eryk Walczak. (2018). "Enhancing central bank communications with behavioural insights". Bank of England Staff Working Paper, 750, Bank of England. https://doi.org/10.2139/ssrn.3233695

Boukherouaa, El Bachir, Ghiath Shabsigh, Khaled AlAjmi, Jose Deodoro, Aquiles Farias, Ebru S. Iskender, Alin T. Mirestean y Rangachary Ravikumar. (2021). "Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance". IMF Departmental Paper, 2021/024, International Monetary Fund. https://doi.org/10.5089/9781589063952.087

Calderaro, Andrea, y Stella Blumfelde. (2022). "Artificial intelligence and EU security: the false promise of digital sovereignty". *European Security*, 31(3), pp. 415-434. https://doi.org/10.1080/09662839.2022.2101885

Carlucci Aiello, Luigia. (2016). "The multifaceted impact of Ada Lovelace in the digital age". *Artificial Intelligence*, 235, pp. 58-62. https://doi.org/10.1016/j.artint.2016.02.003

Carstens, Agustín. (2019). "The new role of central banks". Financial Stability Institute's 20th Anniversary Conference, Basel, 12 March. https://www.bis.org/speeches/sp190314.pdf

Cazzaniga, Mauro, Carlo Pizzinelli, Emma Rockal y Marina Mendes Tavares. (2024). "Exposure to Artificial Intelligence and Occupational Mobility: A Cross-Country Analysis". IMF Working Paper, 116, International Monetary Fund.

Cipollone, Piero. (2024). *Artificial intelligence - a central bank's view*. National Conference of Statistics, Rome, 4 July. https://www.bis.org/review/r240709c.pdf

Congress of the United States of America. (2022). *Algorithmic Accountability Act of 2022*. https://www.congress.gov/bill/117th-congress/senate-bill/3572/text

Doerr, Sebastian, Leonardo Gambacorta y Jose Maria Serena-Garralda. (2021). "Big data and machine learning in central banking". BIS Working Paper, 930, Bank for International Settlements. https://www.bis.org/publ/work930.pdf

European Banking Authority. (2023). *Machine learning for IRB models: Follow-up report from the consultation on the discussion paper on machine learning for IRB models*. https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf

Fernández Bedoya, Ana. (2019). "Inteligencia artificial en los servicios financieros". *Boletín Económico - Banco de España*, 2/2019, Artículos Analíticos. https://repositorio.bde.es/handle/123456789/8448

Fraisse, Henri, y Matthias Laporte. (2022). "Return on investment on artificial intelligence: The case of bank capital requirement". *Journal of Banking & Finance*, 138(106401). https://doi.org/10.1016/j.jbankfin.2022.106401

Grupo de expertos de alto nivel sobre inteligencia artificial. (2018). *Directrices Éticas para una IA Fiable*. Comisión Europea. https://op.europa.eu/es/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Hellwig, Klaus-Peter. (2021). "Predicting Fiscal Crises: A Machine Learning Approach". IMF Working Paper, 150, International Monetary Fund. https://doi.org/10.2139/ssrn.4026328

Japan Cabinet. (2019). *Social Principles of Human-Centric AI*. https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf

Khandani, Amir, Adlar J. Kim y Andrew Lo. (2010). "Consumer Credit Risk Models Via Machine-Learning Algorithms". American Finance Association 2011 Denver Meetings Papers. https://doi.org/10.2139/ssrn.1568864

Lorenz, Philippe, Karine Perset y Jamie Berryhill. (2023). *Initial Policy considerations for Generative Artificial Intelligence*. Organisation for Economic Co-Operation and Development Artificial Intelligence Papers, 1. https://doi.org/10.1787/fae2d1e6-en.

Monetary Authority of Singapore. (2023). *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector*. https://www.mas.gov.sg/-/media/mas/news-and-publications/monographs-and-information-papers/feat-principles-updated-7-feb-19.pdf

Moor, James. (2006). "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years". *AI Magazine*, 27(4), 87. https://doi.org/10.1609/aimag.v27i4.1911

Moreno, Ángel-Iván, Sergio Gorjón and Joaquín Hernáez. (2021). *Computerized text analysis for assessing legal complexity: the practical example of the Circulars of the Banco de España*. 5th International Conference on Public Policy, Barcelona, 5-9 July. https://www.ippapublicpolicy.org/file/paper/60c0beea33381.pdf

Nayak, B., y Nigel Walton. (2024). *Political Economy of Artificial Intelligence: Critical Reflections on Big Data, Economic Development and Data Society*. Palgrave Macmillan.

Organización para la Cooperación y el Desarrollo Económico. (2019). *Recommendation of the Council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Owolabi, Omoshola S., Prince C. Uche, Nathaniel T. Adeniken, Christopher Ihejirika, Riyad Bin Islam y Bishal Jung Thapa Chhetri. (2024). "Ethical Implication of Artificial Intelligence (AI) Adoption in Financial Decision Making". *Computer and Information Science*, 17(1), pp. 49-56. https://doi.org/10.5539/cis.v17n1p49

Parlamento Europeo y Consejo. (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) n.º 2018/858, (UE) n.º 2018/1139 y (UE) n.º 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial). *Diario Oficial de la Unión Europea,* 1689. https://www.boe.es/doue/2024/1689/L00001-00144.pdf

Rebuffi, Sylvestre-Alvise, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles y Timothy Mann. (2021). *Fixing Data Augmentation to Improve Adversarial Robustness*. Thirty-fifth Annual Conference on Neural Information Processing Systems, 6-14, December. https://doi.org/10.48550/arXiv.2103.01946

Riemer, Stiene, Michael Strauß, Ella Rabener, Jeanne Kwong Bickford, Pim Hilbers, Nipun Kalra, Aparna Kapoor, Julian King, Silvio Palumbo, Neil Pardasani, Marc Pauly, Kirsten Rulf y Michael Widowitz. (2023). *A Generative AI Roadmap for Financial Institutions*. Boston Consulting Group. https://www.bcg.com/publications/2023/a-genai-roadmap-for-fis

Rubio, Jennifer, Paolo Barucca, Gerardo Gage, John Arroyo y Raúl Morales-Resendiz. (2020). "Classifying payment patterns with artificial neural networks: An autoencoder approach". *Latin American Journal of Central Banking*, 1(1). https://doi.org/10.1016/j.latcb.2020.100013

Sadok, Hicham, Fadi Sakka y Mohammed El Hadi El Maknouzi. (2022). "Artificial intelligence and bank credit analysis: A review". *Cogent Economics & Finance*, 10(1), 2023262. https://doi.org/10.1080/23322039.2021.2023262

Shabsigh, Ghiath, y El Bachir Boukherouaa. (2023). *Generative Artificial Intelligence in Finance: Risk Considerations*. IMF Fintech Notes, 2023/006, International Monetary Fund. https://doi.org/10.5089/9798400251092.063

Sheehan, Matt. (2023). "China's AI Regulations and How They Get Made". *Horizons Summer,* 2023(24), pp. 108-125. https://www.cirsd.org/files/000/000/010/82/21e461a985f43655b1731b3c1b50cdccb631afaf.pdf.

Shokri, Reza, Marco Stronati, Congzheng Song y Vitaly Shmatikov. (2017). "Membership Inference Attacks Against Machine Learning Models". 2017 IEEE Symposium on Security and Privacy (SP), pp. 3-18. https://doi.org/10.1109/SP.2017.41

Turing, Alan. (1950). "Computing Machinery and Intelligence". *Mind*, 59(236), pp. 433-460. https://doi.org/10.1093/mind/LIX.236.433

United Nations Educational, Scientific and Cultural Organization. (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser y Illia Polosukhin (2017). "Attention Is All You Need". Cornell University. https://arxiv.org/abs/1706.03762

Warwick, K., y Huma Shah. (2015). "Can machines think? A report on Turing test experiments at the Royal Society", *Journal of Experimental & Theoretical Artificial Intelligence*, 28(6), pp. 989–1007. https://doi.org/10.1080/0952813X.2015.1055826

White House, The. (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Cómo citar este documento

Balsategui, Iván, Sergio Gorjón y José Manuel Marqués. (2024). "La inteligencia artificial en el sistema financiero: implicaciones y avances bajo la perspectiva de un banco central". *Revista de Estabilidad Financiera - Banco de España,* 47, otoño. https://doi.org/10.53479/38235