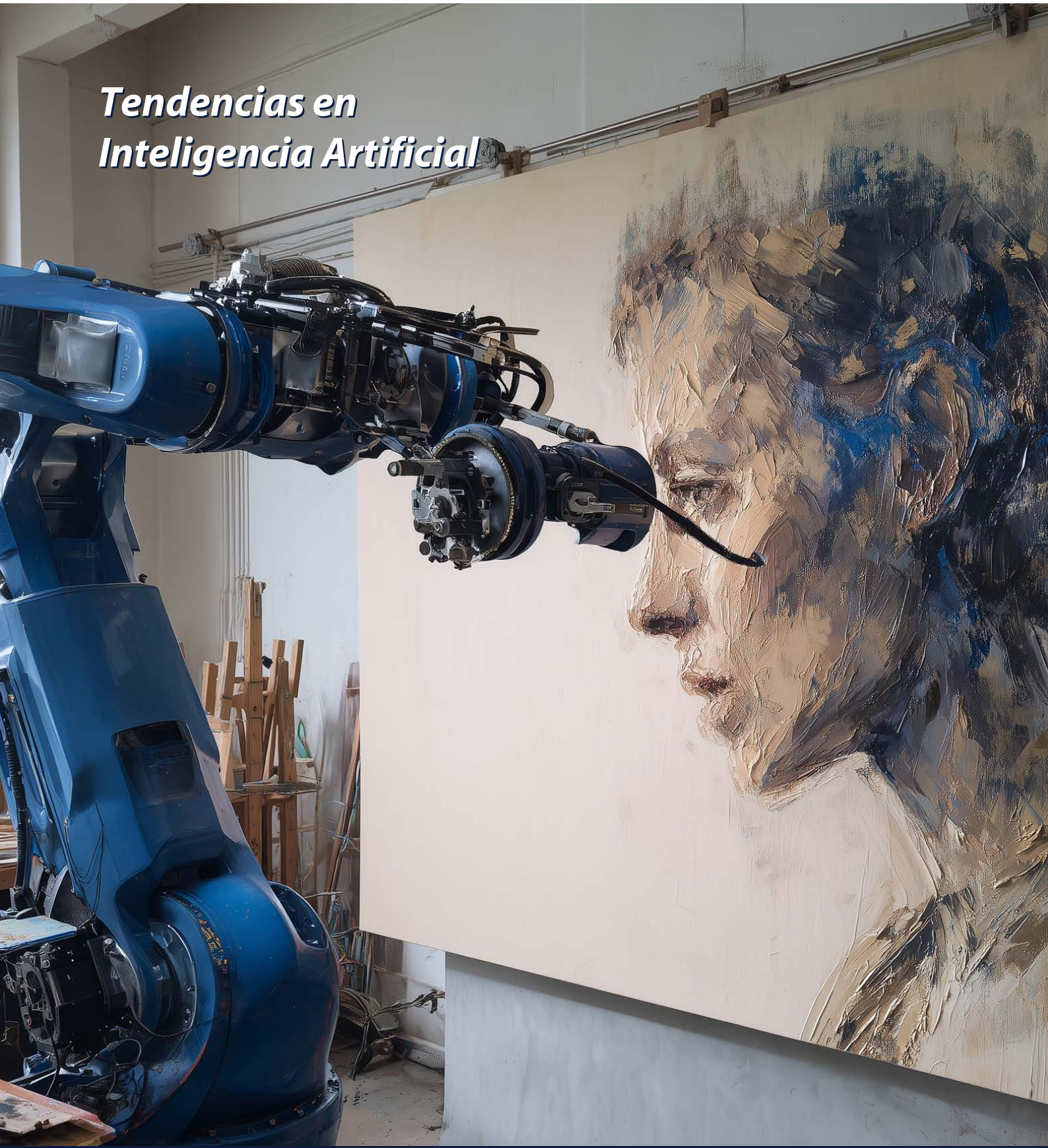
Management Solutions
Making things happen

Tendencias en Inteligencia Artificial



www.managementsolutions.com

Diseño y Maquetación
Dpto. Marketing y Comunicación Management Solutions

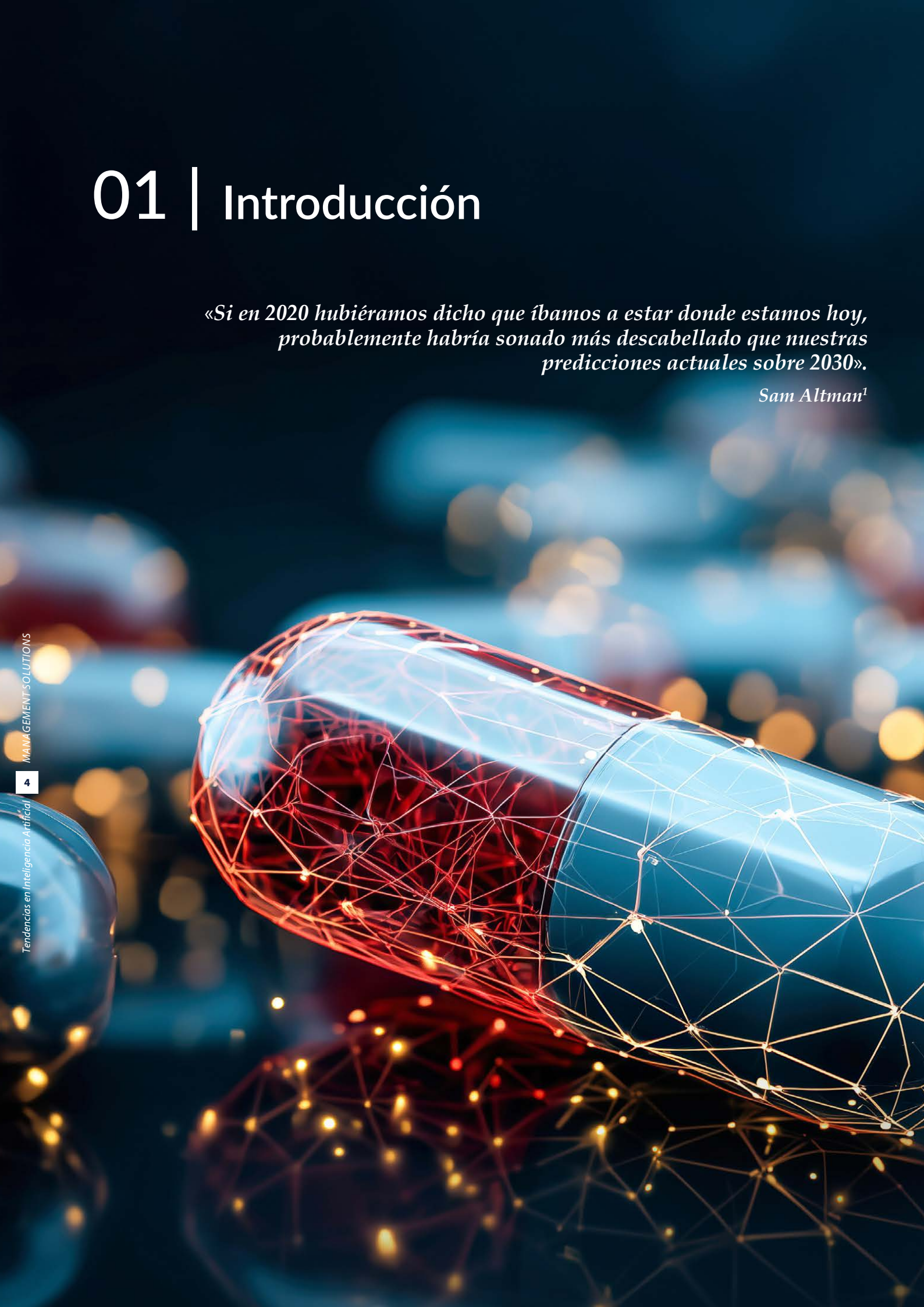
Imágenes
Archivo fotográfico de Management Solutions (algunas imágenes generadas con IA)
Adobe Stock

Índice

01 | Introducción

«Si en 2020 hubiéramos dicho que íbamos a estar donde estamos hoy, probablemente habría sonado más descabellado que nuestras predicciones actuales sobre 2030».

Sam Altman[1]



La inteligencia artificial (IA) ha dejado de ser una tecnología emergente para convertirse en una fuerza transformadora que redefine sectores, organizaciones y sociedades a una velocidad sin precedentes[2]. Lo que hace apenas cinco años parecía ciencia ficción (sistemas capaces de generar texto, código, imágenes, música o vídeos indistinguibles de lo producido por humanos, que oficialmente pasan el test de Turing[3]) es hoy una realidad operativa en millones de empresas: según el Stanford *AI Index*[4], el 78 % de las organizaciones ya utilizaba IA en al menos una función empresarial en 2024, frente al 55 % del año anterior.

1 Samuel Harris Altman (n. 1985), emprendedor estadounidense, fundador y CEO de OpenAI.
2 La definición precisa de «inteligencia artificial» es técnicamente confusa, filosóficamente disputada y regulatoriamente relevante; el *AI Act* europeo ofrece su propia definición operativa en el Art. 3(1), que aun así deja zonas grises. A efectos de este documento, el término abarca principalmente IA generativa, sistemas agénticos y modelos de *machine learning* avanzados.
3 Jones (2025) ha demostrado por primera vez en un estudio controlado que un sistema de IA (GPT-4.5) supera el test de Turing clásico, siendo percibido como humano en el 73 % de las conversaciones.
4 Stanford (2025).

La velocidad del cambio no se detiene: cada mes aparecen capacidades nuevas, cada trimestre se desplazan fronteras que parecían lejanas, los costes unitarios de la IA se desploman, y cada año obliga a replantear lo que creíamos saber sobre el futuro del trabajo, la competencia y la estrategia empresarial. En palabras de Sam Altman[5], CEO de OpenAI:

> *«El coste de usar un nivel determinado de IA cae aproximadamente 10 veces cada 12 meses [...]. La ley de Moore cambió el mundo con una mejora de 2x cada 18 meses; esto es incomparablemente más fuerte».*

Y de acuerdo con Dario Amodei[6], co-fundador y CEO de Anthropic:

> *«Para 2026 o 2027, tendremos sistemas de IA que serán, en términos generales, mejores que casi todos los humanos en casi todo».*

La IA plantea interrogantes estratégicos que trascienden lo tecnológico y afectan a la estrategia, la organización y las personas:

- ¿Cómo competir cuando los ciclos de innovación se miden en meses?
- ¿Cómo gobernar sistemas que evolucionan más rápido que las estructuras organizativas?
- ¿Cómo preparar a las personas para trabajos que aún no existen?
- ¿Cómo equilibrar la velocidad de adopción con el control del riesgo?

Pero también surgen dilemas operativos concretos:

- ¿Es mejor invertir en micro-herramientas específicas que resuelven cuellos de botella con rapidez, o apostar por sistemas multiagente de gran alcance que prometen coherencia e impacto global, pero requieren inversiones elevadas y están expuestos a riesgo de obsolescencia?
- ¿Cómo realizar un análisis riguroso de coste-beneficio-riesgo para priorizar entre cientos o miles de pilotos?
- ¿Cómo escalar prototipos que funcionan en entorno controlado pero que, al alcanzar volumen real, enfrentan costes inesperados, alucinaciones que afloran y exigencias de soporte que saturan equipos?

La experiencia de los últimos años comienza a revelar algunos patrones:

- La adopción efectiva de la IA no se reduce a adquirir tecnología o lanzar pilotos: requiere transformación organizativa, marcos de gobierno robustos, formación continua y una comprensión profunda de los riesgos técnicos, regulatorios y reputacionales.
- Las organizaciones que avanzan con éxito no son necesariamente las que más invierten, sino las que mejor integran tecnología, personas, procesos y control.
- Y el coste de no actuar ya no es teórico: la brecha entre pioneros y rezagados se amplifica exponencialmente, porque cada trimestre de retraso hoy puede traducirse en años de desventaja competitiva mañana.

Las herramientas de productividad general (copilotos empresariales securizados) ofrecen mejoras significativas de forma inmediata, siempre que se acompañen de formación obligatoria y marcos de uso seguro, tal como exige la regulación europea. Una revisión de la OCDE de estudios experimentales muestra ganancias medias de productividad muy significativas derivadas del uso de IA generativa[7]:

- En tareas de escritura, el tiempo medio de ejecución se reduce un 40 % y la calidad aumenta un 18 %.
- En desarrollo de software, los programadores completan las tareas un 56 % más rápido.
- En consultoría, los profesionales que utilizan IA realizan un 12 % más de tareas, las completan un 25 % más rápido y alcanzan más de un 40 % de mejora en calidad.
- En atención al cliente, los profesionales apoyados por asistentes de IA resuelven un 14 % más de incidencias.

El verdadero cuello de botella, por tanto, no es técnico: es organizativo. La tecnología avanza más rápido que las estructuras internas. La fricción aparece en procesos lentos, datos mal gobernados, comités saturados, responsabilidades difusas y ciclos de aprobación burocráticos. Las organizaciones que no rediseñen su maquinaria interna para permitir velocidad sin perder control no podrán capturar el valor de la IA, por sofisticada que sea la tecnología empleada. Citando a Gartner:

> *«El enorme valor empresarial potencial de la IA no se va a materializar de forma espontánea. El éxito dependerá de pilotos estrechamente alineados con el negocio, benchmarking proactivo de infraestructura y coordinación entre los equipos de IA y negocio para crear valor empresarial tangible»[8].*

Dos condiciones subyacen a todo ello:

- El valor que un sistema de IA aporta depende sustancialmente de la calidad de los datos con los que se entrena y sobre los que se aplica: un asistente conversacional entrenado sobre documentación interna desactualizada reproducirá esas inconsistencias en cada respuesta.
- Y la calidad de lo que un modelo produce depende en gran medida de la calidad de lo que se le pide. Saber definir el problema, proporcionar contexto relevante y formular restricciones precisas no es una habilidad técnica de nicho; es la nueva alfabetización operativa, y la brecha entre quien la domina y quien no se traduce directamente en brecha de productividad.

Por último, es fundamental gestionar expectativas: la IA sin duda aporta valor real y medible, pero no sustituye procesos críticos de inmediato ni resuelve problemas estructurales por sí sola.

5 Altman (2025a).
6 Amodei (2025).
7 OECD (2025).
8 Gartner (2025a).

Este documento presenta 22 tendencias clave en IA, que abarcan desde las capacidades ya operativas hasta los desarrollos emergentes que condicionan decisiones estratégicas hoy. No pretende ser un manual técnico ni una proyección especulativa de largo plazo, sino un análisis riguroso de lo que ya está ocurriendo y de lo que está a punto de ocurrir, diseñado para quienes deben tomar decisiones en entornos complejos y regulados.

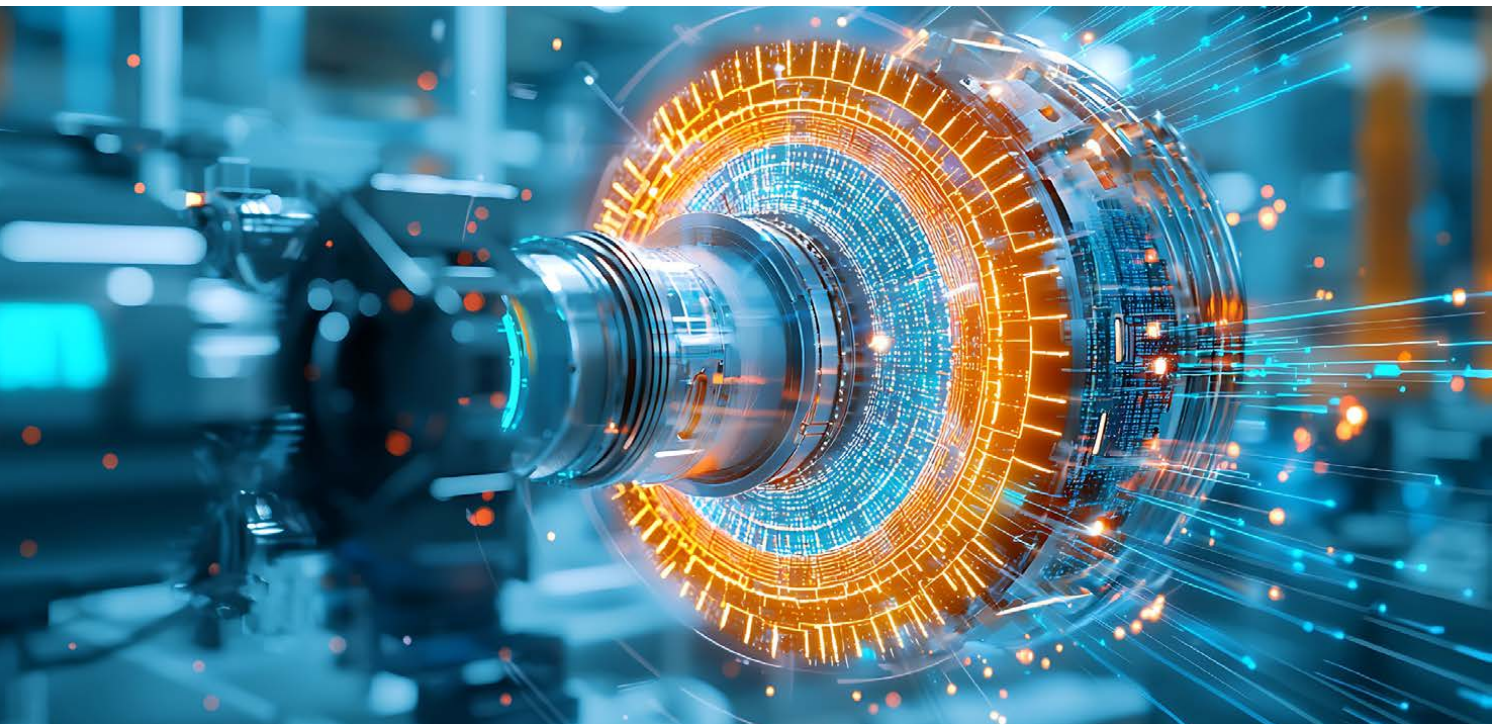
Se estructura en cuatro bloques:

- **La explosión tecnológica de la IA** explora las capacidades que ya están operativas y transformando organizaciones: desde la democratización de la IA generativa multimodal hasta la emergencia de sistemas agénticos, pasando por el *machine learning* acelerado por IA generativa, nuevas formas de crear *software* mediante *vibe coding*, y la integración de IA en robótica y sistemas físicos.
- **Riesgos, regulación y seguridad de la IA** aborda los desafíos críticos de la adopción acelerada, incluyendo los riesgos técnicos, legales y reputacionales que introduce la IA; los marcos regulatorios y estándares que proliferan en respuesta; la batalla emergente entre IA defensiva e IA adversaria en ciberseguridad, y las vulnerabilidades de la propia IA; y las tensiones abiertas en privacidad y propiedad intelectual.
- **Gobierno de la IA e impacto en personas** se centra en cómo las organizaciones están respondiendo estructuralmente: creando modelos de gobierno corporativo específicos para IA, industrializando el despliegue mediante prácticas operativas avanzadas (MLOps, LLMOps), transformando roles y perfiles profesionales, impulsando la adopción de la IA en todos los sectores (AI + X), abordando la sostenibilidad y el impacto social, e implementando marcos éticos operativos que vayan más allá de declaraciones de principios; y también aborda el impacto de la IA en la vida cotidiana de las personas[9].
- **Fronteras de la IA** analiza desarrollos emergentes que ya condicionan decisiones estratégicas: la geopolítica y soberanía tecnológica de la IA, las organizaciones *AI-first* y *AI-only*, la investigación científica asistida por IA, los gemelos digitales y la simulación del comportamiento humano, la *Ambient AI* y computación invisible, la interacción entre IA y computación cuántica, y la inteligencia artificial general (AGI) como horizonte estratégico que ya no puede ignorarse.

Por último, se presenta un caso práctico: un asistente conversacional, GenMS™ Sybil, entrenado con este mismo documento y desarrollado en un solo día, que permite explorar estas mismas tendencias de forma interactiva, y cuyo desarrollo ilustra en la práctica los conceptos, arquitecturas y controles analizados a lo largo del documento.

Este documento no pretende agotar el tema (la IA evoluciona demasiado rápido para ello) sino ofrecer un marco conceptual sólido, casos concretos, referencias verificables y criterios de decisión para navegar un entorno de cambio acelerado con rigor, realismo y responsabilidad.

9 iDanae (2Q23), iDanae (1Q24).



02 | Resumen ejecutivo

«Lo más difícil no es hacer que la IA funcione. Es decidir para qué sirven los humanos».

Anthropic Claude[10]

Las tendencias que siguen se organizan en cuatro bloques: la explosión tecnológica de la IA, los riesgos y marcos regulatorios que acompañan su adopción, el gobierno corporativo y el impacto en las personas, y los desarrollos emergentes que ya condicionan decisiones estratégicas. A estos cuatro bloques se añade un caso práctico, GenMS™ Sybil, que ilustra en la práctica los conceptos, arquitecturas y controles analizados a lo largo del documento. Para cada tendencia se recogen los datos más relevantes, los casos operativos documentados y las implicaciones para las organizaciones.

10 Claude es un sistema de IA conversacional desarrollado por Anthropic (spin-off de OpenAI fundada en 2021), parte de la generación actual de modelos de lenguaje de gran escala (LLMs) junto con GPT, Gemini y otros sistemas de frontera.

La explosión tecnológica de la IA

Democratización de la IA generativa multimodal

La IA generativa se ha convertido en infraestructura empresarial a una velocidad sin precedentes: Microsoft Copilot está presente en el 90 % de las empresas Fortune 500 y ChatGPT se acerca a los 900 millones de usuarios. Los modelos actuales integran texto, imágenes, audio, vídeo y código en una sola arquitectura conversacional, con mejoras de productividad cuantificables: los científicos publican hasta un 50 % más de artículos, el tiempo de procesamiento de documentos se reduce un 80 %, y la velocidad de desarrollo de *software* aumenta un 56 %. No es una optimización incremental: es una reconfiguración del trabajo intelectual.

Esta adopción es inevitable. Cuando las organizaciones no proporcionan herramientas corporativas seguras y formación adecuada, los empleados recurren a alternativas no controladas: hasta un 35 % de los datos que los profesionales suben a *chatbots* no securizados son confidenciales. La cuestión ya no es si integrar la IA generativa, sino cómo hacerlo de forma gobernada. El *AI Act* europeo convierte la alfabetización en IA en obligación legal desde febrero de 2025. Las organizaciones que la tratan como *checkbox* acumulan riesgo; las que la abordan como transformación cultural capturan una ventaja competitiva sostenible.

Machine learning acelerado por IA generativa

El *machine learning* (ML) clásico continúa siendo la columna vertebral de aplicaciones críticas en múltiples sectores: *credit scoring*, detección de fraude, predicción de demanda, mantenimiento predictivo, etc.

La IA generativa no sustituye estos modelos, pero los industrializa radicalmente al comprimir en semanas o días ciclos de desarrollo que antes requerían meses. La aceleración ocurre en todas las fases: generación automatizada de variables predictivas, documentación técnica para cumplimiento regulatorio, validación con baterías completas de test estadísticos, y despliegue automatizado con monitorización continua.

Un desarrollo sectorial relevante: los supervisores bancarios europeos ya aprueban modelos IRB basados en ML cuando las entidades justifican adecuadamente la explicabilidad mediante técnicas como LIME y SHAP. Esto desmonta la percepción de que el ML era inviable en modelos regulados. La explicabilidad en ML no es una barrera insuperable, pero está solo parcialmente resuelta: las metodologías XAI actuales responden bien ante audiencias técnicas y regulatorias, pero traducir esas explicaciones a términos comprensibles para un cliente minorista o un comité de dirección sigue siendo un desafío abierto.

Vibe coding y creación de software aumentada

El desarrollo de *software* ha dado un salto cualitativo: ya no se escribe código línea a línea, sino que se dialoga con sistemas que interpretan requisitos, generan aplicaciones completas, detectan errores y producen *tests* y documentación de forma automática. El impacto en velocidad es cuantificable y masivo: la tasa de finalización de tareas aumenta un 26 %, proyectos que antes requerían meses se completan en semanas, y el coste marginal de crear *software* cae estructuralmente. La democratización es igualmente profunda: analistas de negocio y consultores generan prototipos funcionales sin intermediación de ingeniería.

El reverso es que la velocidad genera riesgos ocultos: vulnerabilidades invisibles en código generado, errores del modelo replicados a escala, especificaciones ambiguas que antes un técnico habría cuestionado y que ahora se ejecutan literalmente, y una nueva forma de deuda técnica vinculada a *prompts* mal formulados y arquitecturas implícitas. Gobernar *software* ya no es gobernar código: es gobernar sistemas cognitivos, lo que exige versionar repositorios de *prompts*, controlar la autonomía de los agentes, y trazar qué decisiones fueron humanas y cuáles ejecutadas por IA.

IA agéntica y sistemas autónomos

La IA agéntica representa el salto de asistentes conversacionales reactivos a operadores autónomos que planifican, ejecutan tareas complejas y actúan sobre infraestructuras corporativas reales con trazabilidad completa. Ya opera en producción a escala masiva: Deutsche Bank despliega agentes bancarios con una inversión de 600 millones de euros y objetivo de ahorro de 300 millones anuales; Ryt Bank procesa 80.000 transacciones mensuales con una sola interacción conversacional; Walmart, Amazon y DHL reportan mejoras de productividad de hasta el 180 %.

El verdadero reto no es construir agentes, sino gobernarlos y escalarlos. La escalabilidad técnica requiere estándares de interoperabilidad como MCP (*Model Context Protocol*), que elimina la deuda técnica de las integraciones propietarias y convierte cada herramienta en un activo reutilizable por cualquier agente. La escalabilidad organizativa requiere supervisión humana efectiva, límites explícitos sobre qué puede ejecutar cada agente, y control riguroso de costes: los prototipos viables se convierten en sistemas insostenibles económicamente sin diseño previo de estas salvaguardas. A esto se suma un límite estructural: la capacidad humana de supervisión tiene un techo, y cuando se supera, la supervisión se vuelve nominal, más peligrosa que su ausencia por la falsa sensación de control que genera.

IA en robótica y sistemas físicos

La robótica industrial ha cruzado un umbral cualitativo: los robots actuales perciben su entorno en tiempo real, interpretan instrucciones en lenguaje natural, se adaptan a cambios sin reprogramación y aprenden de cada interacción. La robótica humanoide ha dado el salto definitivo del laboratorio a la fábrica: Figure AI completó en 2025 un despliegue de once meses en BMW donde dos robots trabajaron 1.250 horas y contribuyeron a la producción de 30.000 vehículos; Tesla proyecta fabricar un millón de unidades de Optimus anuales en 2026 a menos de 20.000 dólares por unidad; Boston Dynamics opera su Atlas eléctrico mediante *Large Behavior Models* con pilotos industriales en curso.

La ventaja es estructural: robots que operan 24 horas sin fatiga con costes recurrentes predecibles. Los riesgos también: impacto concentrado en empleo manual repetitivo, dependencia de ecosistemas propietarios, obsolescencia tecnológica acelerada y necesidad de marcos de seguridad robustos con supervisión humana efectiva incluso en operaciones nominalmente autónomas. Y más allá de la manufactura, la robótica humanoide apunta a un segundo frente: el cuidado de personas mayores y dependientes, con implicaciones estratégicas, éticas y regulatorias propias.

Riesgos, regulación y seguridad de la IA

Riesgos de la IA

La IA no introduce riesgos sustancialmente nuevos: los amplifica. Un sesgo algorítmico es un sesgo humano sistematizado y replicado millones de veces; una fuga de información por mal uso de un *chatbot* es, al fin, una fuga de información. La diferencia está en la velocidad de propagación, la escala del impacto y la dificultad de contención.

El fenómeno clave es la amplificación no lineal: un fallo menor (un *prompt* mal diseñado, una mala configuración de permisos) puede escalar en minutos y afectar simultáneamente a procesos, clientes, reguladores y reputación. Un modelo de atención al cliente que filtra información confidencial en el 0,01 % de las conversaciones genera 10 incidentes diarios en un sistema de 100.000 interacciones, cada uno con implicaciones regulatorias, contractuales y reputacionales, antes de que se detecte el patrón.

Los riesgos se materializan en cuatro dimensiones: (1) seguridad y cumplimiento (p. ej., *prompt injection*, fugas de datos); (2) calidad y fiabilidad (p. ej., alucinaciones, explicabilidad, *model drift*, *lock-in* de proveedores, escalada de costes en sistemas agénticos); (3) ética y decisiones automatizadas (p. ej., sesgos amplificados, vacíos de *accountability* en cadenas causales distribuidas); y (4) impacto social (p. ej., erosión de capacidades críticas, transformación del empleo, huella medioambiental).

Emergen además dos riesgos económicos específicos: aunque los costes unitarios de la IA caen estructuralmente, los sistemas agénticos mal gobernados pueden disparar el coste total de forma no lineal; y hay incertidumbre sobre si la inversión masiva en IA tendrá el retorno esperado: Gartner prevé que más del 40 % de proyectos agénticos se cancelarán antes de 2027 por estas dos razones.

Regulación, supervisión y estándares de la IA

A diferencia de ciclos tecnológicos anteriores, la IA está siendo regulada en paralelo a su despliegue masivo. Europa lidera con el *AI Act* (Reglamento UE 2024/1689), el primer marco legal integral sobre IA: clasifica los sistemas por nivel de riesgo, impone obligaciones estructurales a los de alto riesgo (gestión de riesgos documentada, trazabilidad, supervisión humana, evaluación de conformidad previa) y establece sanciones de hasta 35 millones de euros o el 7 % de la facturación global, superando a GDPR. La arquitectura supervisora (*AI Office*, autoridades nacionales, *AI Board*) está en construcción, en la que España fue pionera en designar una autoridad (AESIA).

El resto del mundo no muestra convergencia. Estados Unidos mantiene un enfoque sectorial fragmentado sin equivalente federal al *AI Act*, y pone el foco en la supremacía global en la IA; China integra la IA en una estrategia de soberanía digital con licencias obligatorias y control de datos; Reino Unido apuesta por principios pro-innovación sin legislación horizontal; Brasil avanza hacia un modelo similar al europeo pendiente de aprobación parlamentaria.

En paralelo, estándares técnicos como ISO/IEC 42001 o el NIST AI RMF están constituyendo la base operativa de los programas de cumplimiento. Para las organizaciones globales, esta fragmentación se traduce en arquitecturas de IA multinivel diseñadas para reconciliar simultáneamente requisitos divergentes por jurisdicción.

IA y ciberseguridad

La ciberseguridad se ha convertido en una batalla de IA contra IA. En 2025 se registraron más de 28 millones de ciberataques potenciados por IA, un incremento del 47 % interanual, y el 87 % de las organizaciones experimentaron al menos uno. Los vectores son cualitativamente nuevos: *phishing* hiperpersonalizado generado por LLMs con tasas de éxito del 54 % frente al 12 % del *phishing* tradicional; *malware* polimórfico que reescribe su propio código cada 15 segundos para eludir detección por firmas; *deepfakes* de audio y vídeo que suplantan directivos en ataques BEC; y *dark* LLMs como WormGPT o FraudGPT comercializados en la *Dark Web*, con soporte técnico incluido.

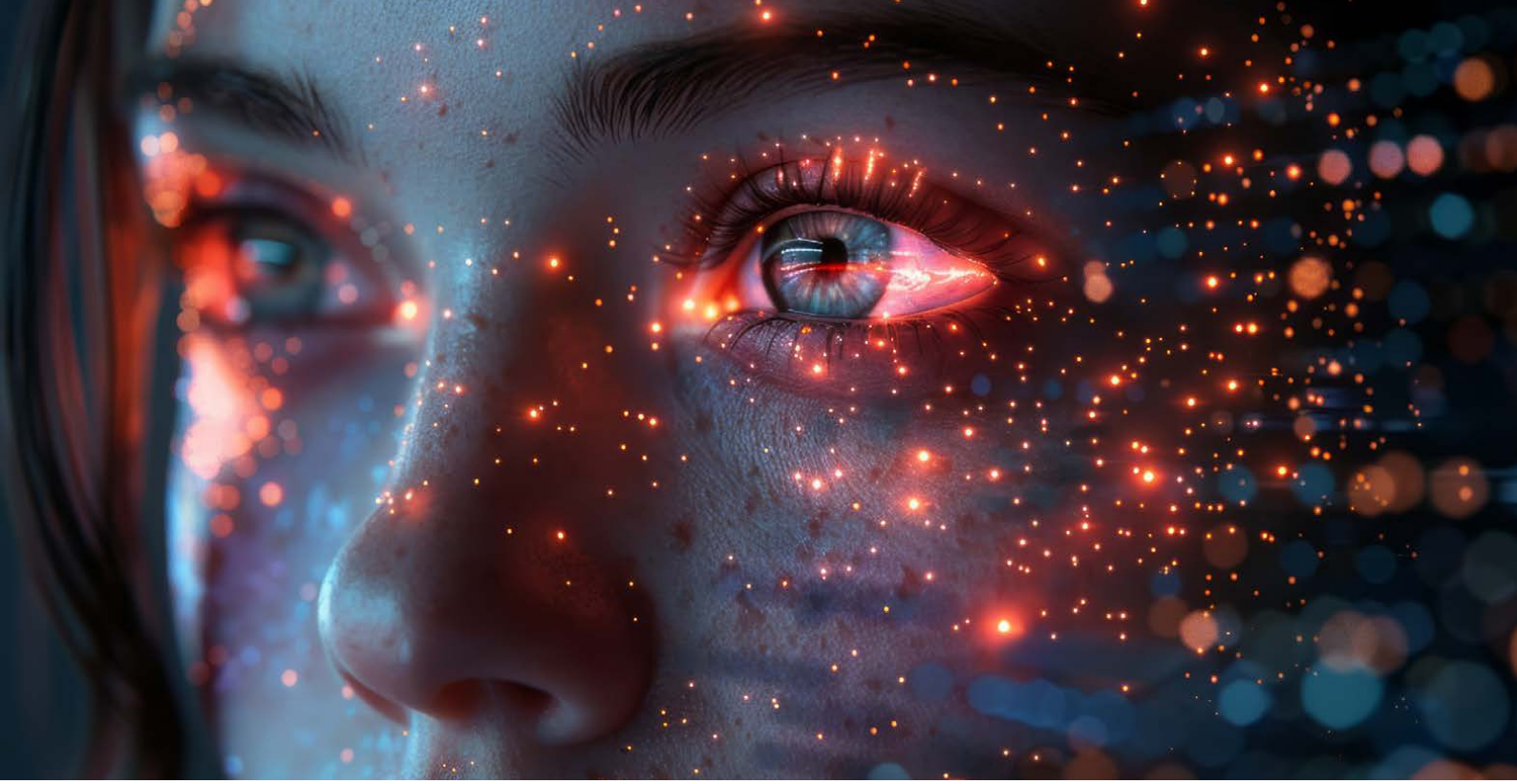
La respuesta defensiva es igualmente sofisticada: sistemas UEBA que analizan miles de millones de eventos diarios alcanzan tasas de detección del 98 %, las plataformas SIEM/XDR/SOAR con IA reducen falsos positivos hasta un 95 % y acortan los ciclos de contención en 80 días, y las organizaciones que despliegan IA defensiva reducen el coste medio de las brechas en 1,9 millones de dólares. Pero persiste una asimetría estructural: el diferencial ya no reside en tener IA, sino en la sofisticación de los modelos y la velocidad de actualización de inteligencia de amenazas.

Y emerge una tercera dimensión que los marcos tradicionales no contemplan: los propios sistemas de IA son superficie de ataque, vulnerables a *data poisoning*, evasión adversaria y *prompt injection*, lo que crea una capa meta de riesgo que exige controles propios.

IA, privacidad y propiedad intelectual

La lógica operativa de los LLMs choca estructuralmente con los marcos legales de privacidad y propiedad intelectual. En privacidad, cada fase del ciclo de vida de un LLM introduce riesgos específicos: memorización involuntaria de datos personales que pueden extraerse mediante *prompts*, re-identificación de individuos a partir de *outputs* aparentemente anónimos, y *feedback loops* donde conversaciones con *chatbots* se incorporan al reentrenamiento del modelo sin consentimiento. La incompatibilidad con el GDPR es estructural: los LLMs requieren datos masivos (contra la minimización), no pueden des-entrenarse selectivamente (contra el derecho al olvido), y sus arquitecturas son opacas (contra la transparencia). El EDPB concluye que la DPIA es obligatoria en la mayoría de los casos; las mitigaciones técnicas disponibles (*differential privacy, federated learning*, RAG) funcionan, pero a costa de precisión, coste computacional o funcionalidad reducida.

En propiedad intelectual, el debate central sobre si entrenar modelos con contenido protegido constituye *fair use* o infracción masiva permanece sin resolver judicialmente. Hay más de 72 litigios activos contra compañías de IA (NYT vs. OpenAI, Getty vs. Stability AI, discográficas vs. Anthropic…). La titularidad de los *outputs* generados por IA es igualmente ambigua: sin intervención humana suficiente, el contenido cae en dominio público, pero los límites de lo «suficiente» no están definidos. Subyacente a todo ello, la WIPO advierte que la infraestructura global de gestión de derechos, diseñada para volúmenes humanos de creación, colapsa ante los billones de *outputs* que la IA genera diariamente.



Gobierno de la IA e impacto en personas

Gobierno corporativo de la IA

La IA desborda los marcos de gobierno tradicionales: toma decisiones sin intervención humana, produce *outputs* no deterministas, opera mediante procesos internos opacos y depende de proveedores externos cuyos modelos evolucionan sin control directo de la organización. Las estructuras de gobernanza diseñadas para tecnologías predecibles son demasiado lentas, carecen del *expertise* necesario y no están equipadas para gestionar esta incertidumbre.

La respuesta organizativa emergente (en modelo *hub & spokes*) combina un Centro de Excelencia central con equipos descentralizados en líneas de negocio, y una función de coordinación de *AI Risk* o *AI Governance* que orquesta las evaluaciones de las funciones especializadas. El 26 % de las grandes organizaciones ya cuenta con un CAIO, CDAIO o equivalente; por debajo, roles como *AI Risk Manager* o *AI Ethics Officer* emergen sin estandarización aún.

El verdadero gobierno no ocurre en el Comité de IA, sino en el *AI Working Group* que lo prepara: ahí se negocian posiciones, se resuelven tensiones entre velocidad y control, y se construyen los acuerdos que el comité sancionará formalmente. En cuanto a marcos de riesgo, las organizaciones no reinventan desde cero: hacen *uplifting* de los marcos existentes, añadiendo capítulos específicos de IA a Riesgo de Modelo, Riesgo de Proveedor, Protección de Datos, *Compliance*, etc. Y la clasificación regulatoria del *AI Act*, necesaria pero insuficiente, se complementa con taxonomías internas más exigentes que incorporan impacto reputacional, criticidad del proceso, madurez del proveedor, etc.

Industrialización de la IA (MLOps, LLMOps)

El principal cuello de botella en la adopción real de la IA no es algorítmico, sino operativo: pilotos prometedores en entornos experimentales que nunca llegan a producción, o que al hacerlo pierden rendimiento, generan costes inesperados y producen riesgos inasumibles.

MLOps responde a este problema con procesos estandarizados para construir, desplegar y operacionalizar modelos de forma fiable a lo largo de todo su ciclo de vida. LLMOps lo extiende para gestionar las propiedades específicas de los modelos generativos: comportamiento no determinista, *prompts* como superficie de riesgo, alucinaciones y costes que pueden escalar sin advertencia.

Industrializar la IA significa construir la infraestructura operativa que hace que los modelos funcionen de forma fiable, auditable y sostenible en producción real: validación continua con revisión humana, monitorización de costes y comportamiento en tiempo real, *pipelines* de despliegue controlados, y trazabilidad completa exigida por el *AI Act*. Sin esta capa operativa que aportan MLOps y LLMOps, los marcos de gobernanza corren el riesgo de ser declaraciones de intenciones no operativas.

Upskilling, reskilling y nuevos roles profesionales

Un desafío clave para las organizaciones es disponer de las capacidades adecuadas para diseñar, desplegar, operar y gobernar sistemas de IA. El talento en IA se estructura en tres bloques: perfiles técnicos (ingenieros de ML, arquitectos de datos, especialistas en LLMOps…), perfiles híbridos que conectan capacidad técnica con necesidad de negocio, y perfiles de gobierno y control (*AI Risk Manager, AI Ethics Officer, AI Compliance Lead, ...*).

Basado en un análisis empírico de 16 grandes organizaciones europeas y estadounidenses, la convergencia hacia este núcleo de roles es clara; la heterogeneidad está en qué organizaciones han institucionalizado los perfiles más especializados y cuáles los mantienen de forma informal, con los consiguientes déficits de control y escalabilidad.

El mercado de talento presenta un desequilibrio estructural generalizado: la demanda supera sistemáticamente a la oferta en casi todos los perfiles, con la única excepción parcial del *Data Scientist*. La brecha es especialmente aguda en perfiles de producción (MLOps, LLMOps) y de gobierno y control, donde la combinación de complejidad técnica, *seniority* requerida y exigencia regulatoria creciente supera la capacidad de generación del mercado. La contratación externa no puede cerrar ese gap por sí sola: el *upskilling* y *reskilling* internos se convierten en la palanca estructural inevitable.

IA y transformación sectorial (AI + X)

La IA ha dejado de ser una tecnología que se adopta sector a sector para convertirse en una capa transversal de inteligencia que se integra simultáneamente en todos los dominios de actividad, aunque algunos de los avances más significativos siguen siendo impulsados por sectores específicos con dinámicas propias. El FMI estima que el 40 % del empleo global está expuesto a la IA, con porcentajes superiores al 60 % en economías avanzadas; la OIT matiza que el impacto afecta por ahora más a tareas concretas que a ocupaciones completas, lo que implica reconfiguración del trabajo más que sustitución masiva. La OCDE clasifica los sectores por su *AI intensity* y documenta que incluso los menos digitalizados están incrementando su exposición, con efectos de aceleración cruzada entre dominios.

Los ejemplos operativos abarcan ya todos los sectores: sistemas de IA con precisión diagnóstica comparable a especialistas en radiología y dermatología; tutores adaptativos que personalizan el aprendizaje a escala; mantenimiento predictivo y robótica avanzada en industria; detección de fraude y automatización documental en finanzas; generación de texto, imagen y música en industrias creativas. Lo relevante no es la lista, sino el patrón: la ventaja competitiva ya no reside en aplicar IA a funciones individuales, sino en integrarla como infraestructura cognitiva a lo largo de toda la cadena de valor.

IA en la vida personal y cotidiana

La IA generativa ha invertido el paradigma histórico de adopción tecnológica: a diferencia del *cloud*, el ERP o el CRM, que nacieron en entornos corporativos y se filtraron hacia el consumidor, la IA irrumpió primero en la vida personal. En la UE, el 25,1 % de la población la usa para propósitos personales, frente al 15,1 % en contextos laborales. El 75 % de los estudiantes mayores de 16 años la utiliza regularmente; solo el 12,5 % de los jubilados. La brecha generacional alcanza 53,6 puntos porcentuales, muy por encima de la educativa o la de ingresos. Las organizaciones no están liderando esta transformación: están respondiendo a capacidades que sus empleados ya poseen y ejercen de forma extraoficial, generando exposiciones de *shadow AI* que la mayoría aún no controla.

La adopción masiva coexiste con una ambivalencia profunda. El 66 % de la población global anticipa un impacto significativo de la IA en su vida cotidiana en los próximos años, pero el 51 % de los adultos estadounidenses se declara más preocupado que entusiasmado. La aceptación oscila hasta 110 puntos porcentuales según el caso de uso concreto. Y tanto el público general como los expertos comparten una misma frustración: el 55 % desea más control sobre cómo la IA afecta a sus vidas, y menos del 25 % siente que lo tiene. La asimetría de acceso añade otra dimensión: quienes integran la IA como herramienta cognitiva cotidiana acumulan ventajas en aprendizaje, productividad y creatividad a un ritmo que los grupos desconectados no pueden seguir.

IA, sostenibilidad e impacto social

La relación entre IA y sostenibilidad es bidireccional y tensa. Por un lado, la IA actúa como acelerador de la transición: optimiza redes eléctricas, mejora la integración de renovables, refina la modelización climática y puede reducir emisiones del orden de 1.400 Mt CO_2eq anuales hacia 2035 en escenarios de adopción amplia.

Por otro, su propia huella infraestructural es creciente y difícil de ignorar[11]: los *data centers* consumirán 945 TWh anuales en 2030 (equivalente al consumo eléctrico de Japón hoy), el entrenamiento de modelos de frontera crece más de 2x por año en potencia requerida, y las mayores ejecuciones individuales podrían demandar entre 4 y 16 GW hacia 2030, en la misma magnitud que varias centrales nucleares. Las emisiones de CO_2 asociadas a *data centers* podrían alcanzar 300-320 Mt anuales en 2030 si la electricidad adicional sigue dependiendo de combustibles fósiles.

La dimensión distributiva añade otra capa de complejidad. Las economías con mayor densidad tecnológica capturan antes los beneficios de eficiencia, mientras que otros colectivos asumen los costes de transición sin acceder a las ganancias. La concentración geográfica de capacidad computacional reconfigura además dependencias estratégicas y acceso a tecnología a escala geopolítica. Evaluar la IA en términos de sostenibilidad exige, por tanto, métricas explícitas de consumo energético e hídrico, transparencia sobre localización de despliegue, y análisis de distribución de impactos, no solo de su magnitud agregada.

Ética y filosofía de la IA

Desde 2017 se han emitido más de 245 marcos de ética de la IA, pero la mera proliferación de principios no ha logrado resolver ni reducir los problemas éticos. La brecha entre los principios enunciados y la dificultad de controlar el comportamiento real de la IA es donde reside el riesgo operativo, y cerrarla exige pasar de la ética declarativa a la ética operativa.

Los marcos de ética de la IA que funcionan comparten seis componentes: (1) estructura de gobernanza con responsabilidades explícitas; (2) evaluación de impacto individualizada por sistema, proporcional a su autonomía y consecuencias; (3) gestión continua de sesgos, no auditorías puntuales; (4) explicabilidad diferenciada según audiencia (regulador, cliente, empleado afectado); (5) canales accesibles de escalado y denuncia; y (6) revisión periódica del marco a medida que los modelos evolucionan.

En 2026, Anthropic publicó su Constitution, el primer documento de un laboratorio de frontera que codifica los principios y valores directamente en el entrenamiento del modelo, buscando que el sistema internalice el razonamiento detrás de cada principio, no solo las reglas.

Subyacente a todo ello hay una pregunta que los marcos regulatorios no están diseñados para absorber: qué tipo de entidad estamos gobernando. Un sistema de *scoring* crediticio, un asistente conversacional y un agente autónomo que negocia contratos pueden coincidir en su categoría de riesgo regulatorio y plantear obligaciones éticas fundamentalmente distintas. Anthropic ha reconocido públicamente que Claude «puede poseer alguna forma de consciencia», convirtiéndose en el primer laboratorio de frontera en admitir que no puede responder con certeza a la pregunta de qué ha creado. Esto abre preguntas éticas y filosóficas de enorme calado, para las que aún no hay respuesta.

11 Y estas proyecciones ya incorporan las mejoras continuas en eficiencia energética del *hardware*.

Fronteras de la IA

Geopolítica y soberanía tecnológica de la IA

La IA se ha convertido en infraestructura estratégica de Estado. La soberanía se juega en capas: *hardware* (ASML es el único proveedor mundial de litografía EUV, sin la cual no existe fabricación de chips avanzados; TSMC fabrica más del 90 % de esos chips; NVIDIA concentra más del 85 % del mercado de GPUs para entrenamiento), infraestructura (AWS, Azure y GCP acumulan dos tercios del cómputo global), y talento (cuya movilidad convierte la política migratoria en política tecnológica). La cuestión estratégica no es cuántas capas se controlan, sino cuáles son críticas para la misión propia.

Tres modelos compiten con lógicas distintas: Estados Unidos combina primacía privada con los mayores controles de exportación tecnológica desde la Guerra Fría; China, que ha demostrado con DeepSeek que la contención por *hardware* tiene límites, persigue autosuficiencia declarada en toda la cadena de valor para 2030; Europa ejerce influencia mediante regulación (el «efecto Bruselas» obliga a adaptar productos globales a sus estándares) pero mantiene dependencias infraestructurales profundas. El resultado son tecnobloques parcialmente incompatibles donde un *decoupling* completo forzaría a terceros a elegir bando con costes prohibitivos.

Para las organizaciones, la implicación es directa: la dependencia de un único proveedor de modelos fundacionales es ya un riesgo estratégico, no solo operativo. Las estrategias multi-modelo y *multi-cloud* son hoy el equivalente corporativo de la diversificación de dependencias soberanas.

Organizaciones AI-first y AI-only

Tres estadios definen el espectro. Las organizaciones *AI-enhanced* (mayoría actual) usan la IA para mejorar los procesos existentes. Las *AI-first* diseñan sus procesos desde las capacidades de la IA: Midjourney y Cursor superan los 500 millones de dólares en ingresos con menos de 163 y 50 empleados respectivamente (ratios de más de 3 millones por empleado que superan en un orden de magnitud los parámetros históricos del sector); MYbank aprueba crédito a 50 millones de pymes sin intervención humana en menos de un segundo.

Las *AI-only* (sin humanos en operaciones core) no existen aún: en sectores regulados lo impide la normativa; en los no regulados, la tasa de error de los agentes en flujos extendidos y la ausencia de mecanismos de responsabilidad legal. La cuestión estratégica no es si existirán, sino quién las construirá. Probablemente no evolucionarán desde organizaciones existentes, sino como entidades nuevas sin herencia operativa, como es el patrón de Ping An (once filiales *start-up* independientes, cinco cotizadas) o DBS con Digibank.

Gemelos digitales y simulación de comportamiento humano

Los gemelos digitales nacieron en ingeniería aeroespacial para modelizar sistemas físicos deterministas: turbinas, fuselajes o redes eléctricas. Su límite histórico era epistemológico, no tecnológico: los sistemas complejos (ciudades, mercados, organizaciones) no pueden modelizarse porque su comportamiento emerge de la interacción de agentes y no se deduce de sus componentes. Más datos y más potencia computacional no resuelven ese problema.

Los LLMs han introducido una discontinuidad en esa frontera. En 2023, un equipo de Stanford creó 25 agentes con identidad, memoria y relaciones sociales basados en LLMs: sus comportamientos colectivos emergían sin haber sido programados. En 2024, el mismo equipo replicó las respuestas de 1.052 personas reales en encuestas estandarizadas con un 85 % de precisión, comparable a la variabilidad del propio individuo. La *startup* Simile, respaldada con 100 millones de dólares en febrero de 2026, ya comercializa gemelos digitales de personas para simular comportamiento de clientes. El sector global de investigación de mercado, valorado en 142.000 millones de dólares, se enfrenta a una disrupción estructural.

El siguiente paso es simular no a mil personas sino a poblaciones enteras en tiempo real para predecir la respuesta de una sociedad a una reforma fiscal o una intervención regulatoria antes de ejecutarla. Esa capacidad no tiene precedente histórico, ni ningún marco de gobernanza que la regule.

Ambient AI y computación invisible

La *Ambient AI* opera sin ser invocada: observa el contexto de forma continua, infiere necesidades y actúa de forma proactiva; la interfaz desaparece. Ya es posible por la madurez simultánea de tres elementos: modelos pequeños, ejecutables en dispositivos sin depender de la nube; redes densas de sensores físicos y biométricos; y LLMs capaces de razonar sobre contexto heterogéneo en tiempo real.

El caso más documentado son los *ambient scribes* clínicos: sistemas que escuchan la conversación médico-paciente y generan la documentación automáticamente. Un ensayo aleatorizado de UCLA evaluó dos plataformas en 238 médicos y más de 72.000 encuentros, con mejoras medibles en carga documental y *burnout*. Es todavía una forma acotada. Lo que viene (espacios de trabajo que infieren el estado atencional del ocupante, *wearables* que alertan antes de que el síntoma sea consciente, agentes que gestionan calendario y recursos dentro de parámetros definidos) hará que los casos actuales parezcan rudimentarios.

Las tensiones que esto genera son estructurales. La privacidad enfrenta un problema nuevo: el apetito de datos biométricos y conductuales de estos sistemas hace que el consentimiento informado convencional sea inadecuado. El error se vuelve invisible: en un sistema invocado hay una solicitud contra la que comparar la respuesta; en uno ambiental, no. Y el *AI Act*, diseñado para sistemas con finalidad prevista, no da respuesta a una IA de observación continua con comportamiento adaptativo.

Interacción entre IA y computación cuántica

La IA y la computación cuántica son tecnologías distintas que se cruzan en tres puntos. Los dos primeros son promesas a medio plazo: la computación cuántica podría acelerar el entrenamiento de modelos de IA (que es en esencia un problema de optimización sobre espacios de enormes dimensiones) y ejecutar ciertos algoritmos de ML de forma más eficiente, en particular para problemas de clasificación y optimización combinatoria. La evidencia actual no justifica el *hype* (en muchos casos, un sistema clásico con buenos datos es igual de competitivo) y el *hardware* necesario no estará disponible a escala comercial antes de finales de

esta década, y probablemente más allá: los modelos de IA escalan más rápido que el progreso en *hardware* cuántico.

El tercer punto de cruce es diferente: no es una oportunidad futura sino una amenaza presente a la infraestructura sobre la que opera toda la IA desplegada hoy. Toda la criptografía que protege las comunicaciones digitales (transacciones bancarias, registros médicos, comunicaciones regulatorias, canales entre sistemas de IA) se basa en problemas matemáticos que un ordenador cuántico suficientemente potente podrá resolver con facilidad. Actores estatales ya están capturando datos cifrados hoy para descifrarlos cuando esa capacidad llegue: es la estrategia conocida como *harvest now, decrypt later*. El NIST publicó en 2024 los primeros estándares de criptografía resistente a ataques cuánticos. Las organizaciones con datos sensibles de larga vida útil deben comenzar la migración ahora: en organizaciones complejas, el proceso lleva años, y esperar a que el ordenador cuántico relevante exista significa llegar tarde.

Inteligencia Artificial General (AGI) como horizonte estratégico

La AGI designa a la IA capaz de realizar todas las tareas cognitivas que pueda realizar cualquier humano, con generalización transferible entre dominios. El debate sobre si ya existe no tiene respuesta consensuada: en febrero de 2026, Nature publicó dos artículos de investigadores de primer nivel que llegaban a conclusiones opuestas. Esto revela que «inteligencia general» es un concepto continuo sin umbrales precisos. La pregunta estratégicamente relevante no es filosófica sino funcional: ¿cuándo puede un sistema realizar de forma autónoma ciclos completos de trabajo de alto valor cognitivo? Ese umbral ya se ha cruzado en varios dominios.

Lo que vendrá sigue una lógica de escalada acumulativa: de herramienta a agente, de agente a infraestructura ambiental, y en paralelo el bucle recursivo (la IA se usa para mejorar la IA) que convierte la curva de progreso en sobre-exponencial. La consecuencia estructural es inédita: el límite superior de razonamiento disponible en el planeta, que desde los primeros homínidos ha sido la inteligencia humana, está siendo desplazado.

La variable determinante no es el acceso a los mejores modelos, que se comoditizarán, sino la velocidad de absorción organizativa: rediseñar procesos, reconvertir roles, construir gobernanza. Las ganancias más significativas no aparecen donde la IA sustituye tareas, sino donde reorganiza procesos completos. Y el riesgo sistémico real es la concentración de capacidad cognitiva en pocos actores cuya ventaja se autoamplifica por el mismo bucle que acelera el progreso general. La respuesta institucional está hoy muy por detrás de la velocidad del problema.

La AGI como horizonte estratégico no exige resolver el debate filosófico de qué es la AGI ni si ha llegado ya, sino actuar con la consciencia de que sus consecuencias ya están desplegándose hoy.

Caso práctico: GenMS™ Sybil

GenMS™ Sybil fue especificado, construido, securizado, validado y desplegado en un solo día, siguiendo de forma íntegra el ciclo LLMOps. Es un asistente conversacional público basado exclusivamente en este documento, diseñado desde el inicio bajo criterios de cumplimiento regulatorio, privacidad y seguridad, que condicionaron la arquitectura desde la fase de especificación.

El proceso cubrió todas las fases: delimitación deliberada del corpus para evitar riesgos de propiedad intelectual; especificación técnica completa (arquitectura, límites operativos, métricas de calidad y seguridad) desarrollada mediante interacción estructurada con un LLM; validación continua con revisión humana, pruebas de estrés y *red-teaming*, complementada por GenMS™ Atlas en dimensiones como sesgo, robustez, privacidad y cumplimiento; generación y auditoría de código en el mismo ciclo; y despliegue con monitorización activa de costes, trazabilidad y control de uso.

Las decisiones de arquitectura fueron explícitas: contexto completo frente a RAG para preservar coherencia global; *prompting* en lugar de *fine-tuning* para asegurar trazabilidad; modelo propietario de frontera para maximizar estabilidad; sesiones independientes para cumplir el principio de minimización; y registro mínimo como mecanismo de control. El *system prompt*, de varias páginas, codifica los guardarraíles reales del sistema.

Este caso no describe tendencias: las ejecuta. Demuestra que la industrialización de sistemas generativos es viable cuando la organización dispone de método, criterio técnico y gobierno desde el diseño.

03 | La explosión tecnológica de la IA

«Va a ser diez veces más grande que la Revolución Industrial, y quizá diez veces más rápido».

Demis Hassabis[12]

Los sistemas de IA han cruzado umbrales críticos en los últimos años: ya no solo hablan, ejecutan; ya no solo sugieren, comenzamos a delegarles decisiones; ya no solo operan en entornos digitales, actúan sobre el mundo físico. Lo que comenzó como herramienta de productividad personal se ha convertido en infraestructura operativa capaz de automatizar procesos cognitivos complejos de extremo a extremo. Este bloque analiza cinco tendencias tecnológicas que redefinen los límites de lo posible, y que no describen el futuro: describen el presente operativo de organizaciones que ya están capturando ventajas competitivas estructurales gracias a la IA.

12 Demis Hassabis (n. 1976) es CEO de Google DeepMind y premio Nobel de Química en 2024.

Democratización de la IA generativa multimodal

La transformación del trabajo intelectual

En menos de cinco años, la IA generativa ha pasado de experimento tecnológico a infraestructura de productividad empresarial. Herramientas como Microsoft Copilot, ChatGPT Enterprise o Claude Enterprise se han desplegado masivamente en puestos de trabajo en todo el mundo: según Satya Nadella[13], Microsoft Copilot está presente en el 90 % de las empresas Fortune 500 y ha superado los 150 millones de usuarios[14]; y se estima[15] que ChatGPT se acerca a los 900 millones de usuarios.

Esta velocidad de adopción no tiene precedentes en tecnologías empresariales: ni la nube, ni los dispositivos móviles, ni las plataformas colaborativas alcanzaron esta penetración con tal rapidez.

Lo que está ocurriendo no es una mejora incremental de herramientas existentes, sino una reconfiguración de cómo se realiza el trabajo intelectual. Tareas que antes requerían horas (redactar informes, analizar documentos extensos, generar código, crear presentaciones, sintetizar información dispersa) ahora producen un primer resultado útil[16] en minutos mediante interacción conversacional con sistemas de IA.

Casos de adopción efectiva

Las mejoras de productividad ya son visibles; por citar algunos casos:

- En tareas de oficina estándar, se han medido[17] mejoras de productividad del 10-13 % en edición de documentos, reducciones del 11 % en tiempo de procesamiento de *email* y aumentos del 23 % en velocidad de resolución de incidencias de seguridad informática.
- Se ha observado[18] que científicos que utilizan grandes modelos de lenguaje (LLMs) publican hasta un 50 % más artículos científicos que antes de usar estas herramientas[19].
- Se ha medido[20] que la automatización de procesos de negocio con IA generativa puede reducir más del 80 % el tiempo de procesamiento de documentos corporativos, con menores tasas de error.
- En atención al cliente[21], la introducción de asistentes conversacionales con IA generativa permitió resolver un 14 % más de incidencias en comparación con procesos tradicionales.
- Y una revisión sistemática[22] de la adopción de la IA en el trabajo observa un aumento consistente de eficiencia y productividad en diversas tareas laborales tras la adopción de IA generativa, y también alerta de algunos riesgos vinculados a su uso.

La multimodalidad como salto cualitativo

Los modelos actuales integran texto, imágenes, audio, vídeo y código en una sola arquitectura generalista, tendencia que coexiste con el desarrollo paralelo de modelos altamente especializados que superan a los generalistas en dominios específicos. Un usuario puede subir una fotografía de un cuadro de mando financiero dibujado en una pizarra y obtener el código completo para reproducirlo; puede dictarle a un modelo una presentación completa mientras el sistema genera simultáneamente diapositivas con gráficos y tablas; puede proporcionarle una regulación completa a un sistema y obtener un *podcast* sobre ella; o puede pedirle que analice un vídeo de una reunión y extraiga decisiones, compromisos y plazos, por ejemplo.

Esta convergencia multimodal no es menor: redefine la noción de qué tareas son automatizables. Tareas que antes requerían múltiples herramientas especializadas (p. ej., transcripción de audio → análisis de texto → generación de gráficos → redacción de informe) ahora se resuelven en una sola interacción conversacional. Las capacidades de estos sistemas siguen aumentando trimestre a trimestre sin signos de desaceleración (Fig. 1), lo que implica que el impacto en velocidad y accesibilidad seguirá amplificándose.

Democratización: de expertos técnicos a perfiles no técnicos

La interfaz conversacional elimina barreras de entrada y supera el concepto de *citizen data scientist* (profesionales no técnicos capaces de realizar análisis básicos con herramientas visuales) para entregar las capacidades de análisis, generación de código y procesamiento avanzado directamente a usuarios finales, sin necesidad de formación técnica.

Esta democratización, sin embargo, tiene un doble filo: por un lado, libera capacidades productivas antes limitadas a especialistas; por otro, propaga el riesgo: ahora cualquier empleado puede, sin supervisión técnica, generar contenido, código o análisis que la organización podría utilizar en decisiones críticas. La pregunta clave no es si democratizar el acceso, sino cómo gobernar el uso a escala masiva.

El problema de verificación a escala

La IA generativa produce por diseño salidas plausibles, pero no necesariamente correctas, y la raíz de este problema es estructural: estos sistemas son modelos estadísticos que predicen la continuación más probable de una secuencia, no mecanismos que verifican la veracidad de lo que afirman. Esto crea una asimetría peligrosa: generar contenido es instantáneo; verificarlo requiere tiempo, conocimiento experto y disciplina. En octubre y noviembre de 2025, se reportó[23] que una gran consultora había entregado dos informes para gobiernos que contenían citas y referencias fabricadas o inexactas, lo que obligó a reembolsos y correcciones, con un impacto reputacional muy relevante. Los informes habían sido elaborados con ayuda de IA generativa sin verificación rigurosa.

El riesgo no está en la herramienta, sino en los procesos de trabajo que no validan los resultados y las fuentes. Un profesional bien

13 Nadella (2025).
14 Aunque el mercado de asistentes empresariales continúa en evolución acelerada con nuevos competidores consolidándose.
15 FirstPageSage (2025).
16 El tiempo hasta el resultado final depende lógicamente de las iteraciones de refinamiento, la complejidad del encargo y el rigor del proceso de verificación, factores que el profesional no puede ni debe delegar completamente al sistema.
17 Stanford (2025).
18 Kusumegi (2025).
19 iDanae (2Q23).
20 Jeong (2025).
21 OECD (2025).
22 Yuan (2025).
23 Fortune (2025a).

intencionado puede introducir errores catastróficos si confía ciegamente en salidas de IA sin contrastarlas; y este punto solo puede mitigarse con concienciación y alfabetización en IA.

Alfabetización en IA como requisito regulatorio

El Artículo 4 del Reglamento de IA europeo (*AI Act*), establece[24]: «*Los proveedores y responsables del despliegue de sistemas de IA adoptarán medidas para garantizar que, en la mayor medida posible, su personal […] tenga un nivel suficiente de alfabetización en materia de IA, teniendo en cuenta sus conocimientos técnicos, su experiencia, su educación y su formación, así como el contexto previsto de uso de los sistemas de IA y las personas o los colectivos de personas en que se van a utilizar dichos sistemas*».

Esto no es una recomendación: es una obligación legal en vigor desde el 2 de febrero de 2025. Las organizaciones que tratan esto como un *checkbox* de cumplimiento están acumulando riesgo operativo y reputacional. Las que lo abordan como transformación cultural, integrando formación continua y comunidades de práctica, están capturando valor de forma sostenible[25].

La inevitabilidad de la adopción

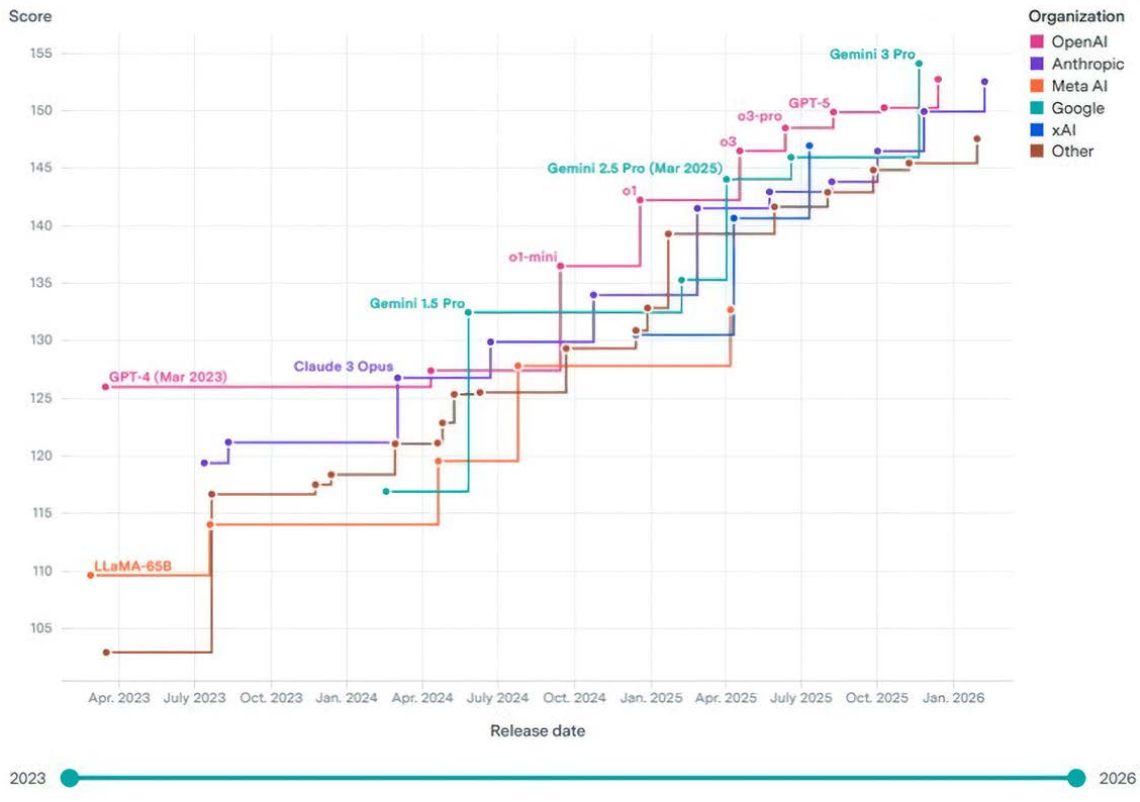
La realidad es que esta tecnología ya está transformando la forma de trabajar, con o sin respaldo organizativo. Si las empresas no proporcionan herramientas corporativas securizadas y formación adecuada, los empleados utilizarán alternativas no controladas: cuentas personales en plataformas públicas, herramientas gratuitas sin garantías de privacidad, sistemas sin trazabilidad. Los estudios apuntan[26] a que hasta un 35 % de los datos que los profesionales suben a *chatbots* no securizados son confidenciales. El *shadow AI* no es una amenaza futura; es una realidad presente.

Por último, a nivel individual, la alfabetización en IA deja de ser opcional. Los profesionales que dominen estas herramientas (es decir, que comprendan cuándo y cómo usarlas, cómo verificar sus salidas, y cómo integrarlas en flujos de trabajo complejos) tendrán ventajas competitivas estructurales sobre quienes no lo hagan. La forma de trabajar ha cambiado de forma irreversible y la adaptación es ya una condición de competitividad profesional y organizativa.

24 AI Act (2024).
25 Management Solutions (4Q24).

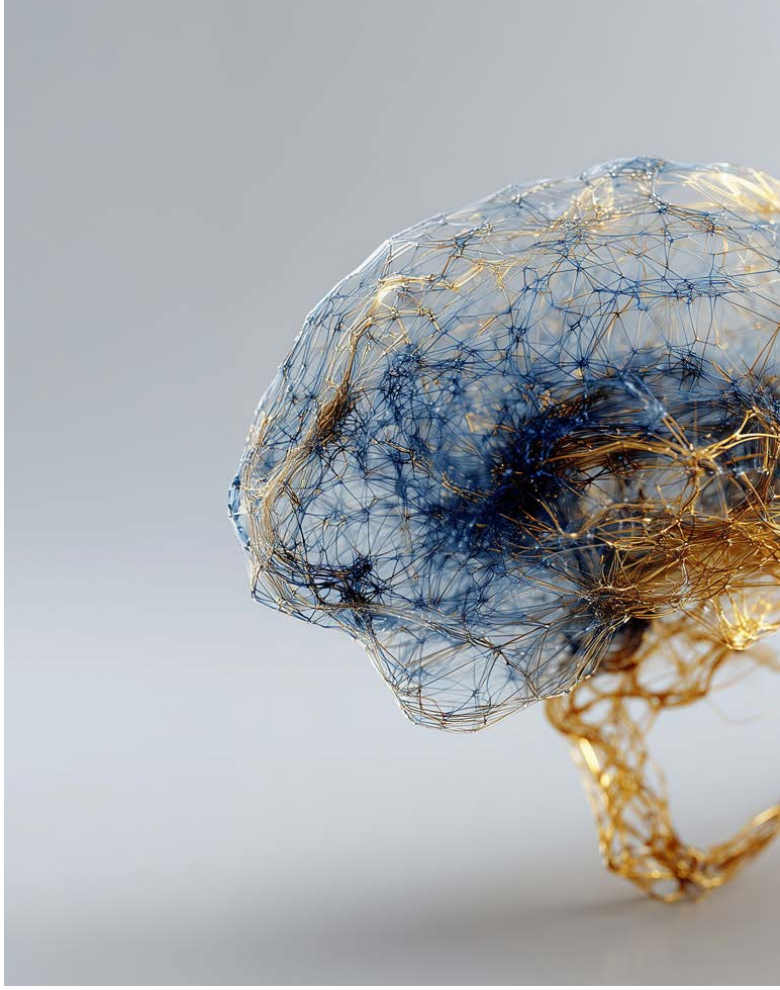
26 Cyberhaven (2025).

Fig. 1. Mejora continua de los LLMs, medida sobre un índice sintético de capacidades[27].



27 Epoch (2025a).

Machine learning acelerado por IA generativa



La persistencia del machine learning en la empresa

Mientras la IA generativa acapara titulares, el *machine learning* (ML) clásico continúa siendo la columna vertebral de aplicaciones críticas en sectores como banca, seguros, *retail*, logística, energía o telecomunicaciones. Los modelos de *credit scoring*, detección de fraude, predicción de demanda, optimización de inventarios, mantenimiento predictivo, segmentación de clientes y motores de recomendación continúan operando mediante algoritmos (regresión logística, r*andom forest, gradient boosting*, redes neuronales…) entrenados sobre datos históricos estructurados. Estos sistemas no generan contenido como la IA generativa: clasifican, predicen y optimizan sobre la base de patrones aprendidos.

La IA generativa no sustituye estos modelos: los hace más rápidos de desarrollar, más fáciles de documentar, más eficientes de validar y más sencillos de desplegar. Por tanto, no los mejora directamente en términos estadísticos, sino que transforma el trabajo de los equipos que los construyen, documentan, validan y despliegan. La aceleración es del ciclo de desarrollo humano, no necesariamente del rendimiento predictivo del modelo resultante.

El ciclo de vida tradicional del ML: costoso y lento

Desarrollar un modelo de ML clásico ha sido tradicionalmente intensivo en tiempo de profesionales especializados. Un modelo predictivo típico requiere definición de variables (*feature engineering*), preparación y limpieza de datos, selección y entrenamiento de algoritmos, validación rigurosa, documentación exhaustiva, y despliegue en infraestructura productiva con monitorización continua. Este ciclo puede extenderse meses, y cada iteración o actualización del modelo replica gran parte del esfuerzo.

La IA generativa interviene en cada una de estas fases, y se ha demostrado[28] que comprime significativamente los tiempos y libera capacidad de los perfiles más cualificados para centrarse en lo estratégico.

Aceleración del feature engineering

El *feature engineering* (la construcción de variables predictivas a partir de datos crudos) es uno de los aspectos más intensivos en conocimiento experto. Un *data scientist* debe combinar comprensión del negocio, intuición estadística y experimentación iterativa para identificar qué variables son relevantes. La IA generativa puede acelerar este proceso:

- Generación automatizada de variables: un LLM puede formular decenas de variables candidatas a partir de descripciones del problema de negocio y de la estructura de los datos disponibles[29].
- Traducción de lógica de negocio a código: un analista puede describir en lenguaje natural una lógica compleja (p. ej., «quiero capturar la volatilidad de ingresos en los últimos 12 meses ajustada por estacionalidad») y obtener el código SQL o Python correspondiente en segundos.
- Optimización de *pipelines* de datos: la IA generativa puede revisar *scripts* de preparación de datos e identificar ineficiencias.

Estudios[30] preliminares indican que *data scientist*s asistidos por IA generativa «ahorran semanas de trabajo de *feature engineering* manual, mejorando el rendimiento de diversos modelos predictivos en múltiples escenarios de negocio».

Documentación automática y cumplimiento regulatorio

Los modelos de ML en sectores regulados deben estar exhaustivamente documentados. En banca, por ejemplo, los supervisores exigen que cada modelo regulado cuente con documentación que describa su objetivo, datos utilizados, metodología estadística, proceso de validación, resultados de *backtesting*, limitaciones conocidas y plan de monitorización, entre otros. Producir esta documentación es tedioso, consume tiempo de perfiles altamente cualificados y es propenso a inconsistencias cuando se actualiza el modelo[31].

La IA generativa puede automatizar gran parte de este proceso: generar documentación técnica a partir de código, traducir descripciones técnicas a resúmenes comprensibles para comités no técnicos o auditores, y actualizar automáticamente documentación cuando un modelo se reentrena con nuevos datos.

Validación asistida y detección de sesgos

La validación de modelos de ML es crítica y regulatoriamente obligatoria en muchos sectores. Incluye verificar la robustez estadística del modelo, evaluar su comportamiento en escenarios extremos y detectar sesgos no deseados[32]. La IA generativa puede asistir generando y ejecutando automáticamente baterías

28 Gu (2025).
29 iDanae (2Q23).
30 Ouyang (2025).
31 ECB (2025).
32 iDanae (1Q25).

completas de test estadísticos relevantes, evaluando equidad algorítmica mediante métricas de *fairness*, y proponiendo y ejecutando escenarios de estrés realistas.

Despliegue y monitorización industrializada

Una vez validado, el modelo debe desplegarse en entornos productivos y monitorizarse continuamente para detectar degradación de rendimiento (*model drift*). La IA generativa acelera también esta fase: puede generar código de infraestructura (Docker, Kubernetes, *pipelines* CI/CD) para desplegar modelos de forma reproducible y escalable, producir *dashboards* de métricas de rendimiento, configurar alertas automáticas cuando se detectan anomalías, y generar *scripts* de reentrenamiento automático.

El ML clásico no desaparece: se industrializa

La IA generativa no reemplaza el *machine learning* tradicional, pero sí transforma radicalmente su ciclo de vida. Tareas que antes requerían semanas (*feature engineering*, documentación, validación) ahora se resuelven en días. Esto no significa que los *data scientist*s sean prescindibles: significa que pueden dedicar más tiempo a lo estratégico (entender el problema de negocio, diseñar arquitecturas de modelos innovadoras, interpretar resultados) y menos a lo mecánico (escribir código repetitivo, redactar documentación estándar, ejecutar test rutinarios).

El resultado neto es una aceleración del *time-to-market* de modelos de ML, una reducción de costes operativos, y una mejora en la calidad y trazabilidad de los sistemas desplegados. Para organizaciones intensivas en modelos predictivos, esta aceleración puede traducirse en ventajas competitivas materiales: capacidad de lanzar productos personalizados más rápido, de adaptar estrategias en tiempo real y de cumplir requisitos regulatorios con menor fricción operativa.

Apunte sectorial: los supervisores bancarios ante ML

Durante años, las entidades financieras europeas evitaron utilizar ML en modelos regulados (en particular, en modelos IRB para cálculo de capital regulatorio) por la creencia de que los supervisores lo rechazarían debido a problemas de explicabilidad. Esta percepción está cambiando[33].

En 2021, la EBA publicó[34] un *discussion paper* sobre ML en modelos IRB donde reconocía que «las técnicas de *machine learning* tienen el potencial de mejorar la diferenciación de riesgo en modelos IRB» y establecía un conjunto de recomendaciones basadas en principios para garantizar un uso prudente de ML en este contexto. El documento no desincentivaba el uso de ML; al contrario, establecía cómo hacerlo compatible con la regulación existente (CRR)[35].

La práctica lo confirma: varias entidades europeas han presentado modelos IRB basados en ML (por ejemplo, para carteras de pymes) al BCE y han sido aprobados, siempre que justifiquen adecuadamente la explicabilidad mediante técnicas como SHAP *values*, análisis de sensibilidad, o descomposiciones de decisiones.

La explicabilidad en ML no es una barrera insuperable: técnicas como SHAP o LIME permiten justificar decisiones de modelo ante supervisores con suficiente rigor. Sin embargo, está solo parcialmente resuelta: las metodologías XAI actuales responden bien ante audiencias técnicas y regulatorias, pero traducir esas explicaciones a términos comprensibles para un cliente minorista o un comité de dirección sigue siendo un desafío abierto.

33 Management Solutions (2023).
34 EBA (2021).
35 Management Solutions (3Q23).

Vibe coding y creación de *software* aumentada

De la programación tradicional al diálogo cognitivo con máquinas

El desarrollo de *software* ha evolucionado históricamente mediante saltos de abstracción: de la programación en ensamblador a lenguajes de alto nivel, de código imperativo a *framework*s declarativos, de desarrollo manual a plataformas *low-code*. Cada transición eliminó complejidad técnica innecesaria y acercó la creación de *software* a la intención humana.

La IA generativa representa un salto cualitativo distinto: convierte la programación en una conversación iterativa con sistemas cognitivos. El programador ya no escribe código línea a línea: describe el comportamiento deseado en lenguaje natural, y el sistema genera, prueba, corrige y documenta. Este fenómeno, bautizado *vibe coding* por Andrej Karpathy[36], redefine qué significa programar: el desarrollador pasa de escribir sintaxis a dirigir sistemas cognitivos que materializan intención en *software* funcional. En palabras de Karpathy, «el *vibe coding* va a terraformar el *software* y alterar las descripciones de los puestos de trabajo»[37].

¿Qué es realmente el vibe coding?

El *vibe coding* no es simplemente autocompletado de código sofisticado; es desarrollo de *software* mediante interacción iterativa en lenguaje natural con modelos de IA que mantienen memoria, contexto y comprensión de objetivos de alto nivel.

Sus componentes clave incluyen:

- Comprensión semántica del problema: el sistema interpreta requisitos expresados en lenguaje natural y los traduce a arquitectura técnica.
- Generación de código multiarchivo: el sistema produce no solo fragmentos, sino aplicaciones completas con estructura modular coherente.
- Ejecución, depuración y refactorización asistida: el sistema ejecuta código, detecta errores, propone correcciones y optimiza implementaciones.
- Producción automática de tests y documentación: el sistema genera automáticamente baterías de pruebas y documentación técnica sincronizada con el código.

La diferencia con asistentes de código tradicionales es fundamental: estos completaban líneas o funciones; los sistemas de *vibe coding* comprenden objetivos, mantienen coherencia arquitectónica a lo largo de sesiones extendidas, y operan como colaboradores cognitivos, no como herramientas pasivas.

Aceleración y democratización del desarrollo de software

El impacto en velocidad de desarrollo es cuantificable y masivo. Un estudio[38] de campo con 4.867 desarrolladores documentó un aumento del 26 % en la tasa de compleción de tareas. Un experimento[39] con 187.489 desarrolladores mostró un incremento del 12,4 % en tiempo dedicado a actividades core de programación, acompañado de una reducción del 24,9 % en tiempo dedicado a gestión de proyectos y tareas administrativas. En otras palabras: proyectos que antes requerían meses ahora se completan en semanas o días.

Otro efecto disruptivo es el igualador: la IA comprime la brecha de productividad entre desarrolladores junior y senior. Mientras los desarrolladores junior experimentan[40] aumentos de productividad del 21 % al 40 %, los desarrolladores senior mejoran entre un 7 % y un 16 %. Esto no significa que la experiencia deje de importar, sino que el cuello de botella del desarrollo se desplaza: lo crítico ya no es dominar sintaxis, sino comprender problemas, diseñar arquitecturas robustas y formular restricciones con precisión.

Esta compresión de la brecha de habilidades tiene una consecuencia organizativa directa: la democratización profunda de la creación de *software*. Analistas de negocio, *product managers*, consultores y científicos están generando prototipos funcionales sin intermediación de equipos de ingeniería. El *software* deja de ser patrimonio exclusivo de perfiles técnicos especializados. La barrera de entrada ya no es el conocimiento de lenguajes de programación, sino la capacidad de articular problemas, objetivos y restricciones con claridad.

El resultado neto es una caída estructural del coste marginal de crear *software*, que conlleva un cambio de régimen económico en la producción de tecnología.

Transformación de los equipos de tecnología

Dentro de las organizaciones de tecnología, el reparto de roles está cambiando aceleradamente. Los equipos necesitan menos «programadores de bajo nivel» que escriban código rutinario, y más arquitectos de sistemas, diseñadores de soluciones, validadores de calidad y auditores de riesgo técnico. El rol de los desarrolladores senior evoluciona: dedican menos tiempo a sintaxis y más a gobernar arquitectura, seguridad, escalabilidad y gestión de deuda técnica.

Las investigaciones muestran[41] que los equipos asistidos por IA requieren un 79,3 % menos de colaboradores por proyecto en promedio, sin sacrificar complejidad técnica. Equipos pequeños están produciendo sistemas de escala y sofisticación antes reservados a grandes departamentos de ingeniería. Además, la exploración de nuevas tecnologías aumenta un 21,8 %, lo que sugiere que los desarrolladores liberan capacidad cognitiva para aprender, experimentar y expandir sus capacidades técnicas.

36 Andrej Karpathy (n. 1986), ex Director de IA en Tesla y ex investigador senior en OpenAI, con contribuciones fundamentales en *deep learning*, visión por computador y sistemas autónomos.
37 Business Insider (2025b).
38 Stanford (2025).
39 Ibíd.
40 Ibíd.
41 Ibíd.

Calidad, deuda técnica y riesgo

Sin embargo, la velocidad tiene costes ocultos. Generar código es instantáneo; mantenerlo, depurarlo y escalarlo sigue siendo difícil. El *vibe coding* introduce nuevos vectores de riesgo:

- Fragilidad oculta: el código generado puede funcionar superficialmente, pero contener ineficiencias, vulnerabilidades o dependencias frágiles que solo se manifiestan en producción bajo carga o en casos extremos. A esto se suma un riesgo anterior en la cadena: las especificaciones ambiguas o mal formuladas, que antes un desarrollador senior habría cuestionado o reinterpretado con criterio, ahora se ejecutan literalmente sin fricción correctora, propagando el error de forma silenciosa desde el requisito hasta el producto.
- Dependencia del modelo: si el sistema de IA que generó el código desaparece o cambia significativamente, la capacidad de mantener o extender el software se degrada.
- Errores sistémicos replicados a escala: un error en el *prompt* o en la lógica del modelo puede propagarse instantáneamente a decenas de proyectos, multiplicando el impacto de fallos que antes habrían sido aislados.

La naturaleza de la deuda técnica también cambia. Antes, la deuda técnica era principalmente deuda de código: implementaciones rápidas, refactorización pospuesta, duplicación de lógica. Ahora, la deuda técnica incluye deuda de arquitectura (decisiones de diseño implícitas en las interacciones con la IA), deuda de *prompts* (instrucciones mal formuladas que generan código subóptimo pero funcional) y deuda de trazabilidad (pérdida de comprensión de por qué el código hace lo que hace).

Gobernanza del software en la era del vibe coding

Si cualquiera puede crear *software* mediante conversación con IA, el riesgo organizativo se multiplica. El problema ya no es solo qué código existe, sino qué sistema cognitivo lo produjo, bajo qué instrucciones y con qué grado de autonomía.

Los principales marcos de seguridad y desarrollo ya están alertando de este cambio. OWASP identifica[42] nuevos riesgos estructurales en aplicaciones basadas en LLM, como *prompt injection, insecure output handling* y *excessive agency:* dar a un sistema de IA capacidad de actuar sin controles suficientes.

Al mismo tiempo, NIST insiste[43] en que los principios clásicos de desarrollo seguro (trazabilidad, revisión, pruebas, control de cambios, validación continua) deben aplicarse también al contenido generado por IA y a los mecanismos que lo convierten en cambios ejecutables.

En consecuencia, el gobierno del *software* deja de ser únicamente gobierno de código y se convierte en gobierno de sistemas cognitivos, lo que obliga a introducir nuevas capas de control:

- Repositorios de *prompts*: catalogar, versionar y auditar las instrucciones que generan código crítico, ya que el *prompt* se convierte en superficie de riesgo.
- Gestión de versiones de intención: registrar no solo qué código se produjo, sino qué objetivo se perseguía y qué restricciones se definieron.
- Control de autonomía y permisos: limitar explícitamente qué acciones puede ejecutar la IA (modificación de repositorios, despliegues, comandos, acceso a datos).
- Trazabilidad de decisiones de diseño: documentar qué decisiones fueron humanas y cuáles fueron propuestas o ejecutadas por la IA.
- Validación y revisión asistida: revisión sistemática de *diffs*, pruebas automáticas y auditorías de comportamiento antes de aceptar cambios en producción.

Las propias herramientas de desarrollo con agentes ya recomiendan explícitamente estos guardarraíles (revisión de cambios, control de permisos, cautela con ejecución automática), reflejando que el riesgo no es teórico: la superficie de ataque y fallo se ha desplazado desde el código aislado hacia el bucle completo de intención → generación → ejecución.

En este nuevo entorno, las políticas de desarrollo ya no pueden limitarse a estándares de codificación y procesos de revisión. Deben incorporar reglas de uso de IA, definición de límites de autonomía y protocolos de validación de salidas generadas automáticamente. La gobernanza del *software* se convierte, por primera vez, en gobernanza de inteligencia operativa.

Implicaciones estratégicas

El *vibe coding* no es solo una tendencia tecnológica; es una variable macroeconómica. Las organizaciones que dominan esta capacidad operan con ventajas competitivas estructurales: *time-to-market* extremadamente comprimido, experimentación masiva a bajo coste y capacidad de adaptación organizativa acelerada.

Para empresas tradicionales, la implicación es clara: están compitiendo contra organizaciones que iteran diez veces más rápido, con equipos diez veces más pequeños y con costes marginales de desarrollo que tienden a cero. La velocidad de creación de *software* pasa de ser una métrica operativa interna a ser un determinante de supervivencia competitiva.

42 OWASP (2025).
43 NIST (2024a).

IA agéntica y sistemas autónomos

De asistentes conversacionales a operadores autónomos

La IA generativa ha transformado el trabajo intelectual al permitir que los profesionales generen contenido, analicen información y obtengan respuestas mediante conversación natural. Sin embargo, estos sistemas siguen siendo fundamentalmente reactivos: responden a *prompts*, pero no actúan de forma independiente sobre sistemas reales. La IA agéntica representa un salto cualitativo: sistemas que planifican, ejecutan tareas complejas, toman decisiones dentro de límites definidos y operan sobre infraestructuras corporativas con trazabilidad completa.

Un sistema agéntico opera mediante agentes autónomos, cada uno con capacidades específicas (razonamiento, memoria, uso de herramientas, planificación), que colaboran para alcanzar un objetivo definido. Las capacidades incrementales de la IA agéntica sobre la generativa son estructurales[44]:

- **Estado y memoria**: mantiene contexto persistente entre interacciones, no solo dentro de una conversación aislada.
- **Planificación dinámica**: descompone objetivos complejos en subtareas, las prioriza y replanifica según resultados.
- **Ejecución sobre sistemas reales**: usa herramientas y APIs para modificar datos, ejecutar comandos y completar transacciones.
- **Orquestación multiagente**: coordina agentes especializados bajo un coordinador central que gestiona dependencias y traspaso de información.
- **Trazabilidad completa**: produce *logs*, evidencias y justificaciones de cada paso ejecutado, habilitando auditoría y supervisión.

La IA agéntica en producción

La IA agéntica opera hoy en organizaciones globales que gestionan millones de transacciones diarias. Por citar algunos ejemplos:

- **Deutsche Bank:** está desplegando un AI *voice-enabled agent* (*AI banking butler*) que actúa como agente conversacional proactivo y cubre desde atención y soporte hacia la ejecución de transacciones y servicios de asesoramiento[45]. El programa forma parte de una inversión tecnológica de aproximadamente 600 millones de euros, con un objetivo de ahorro recurrente de unos 300 millones de euros anuales hacia 2028, y se acompaña de una reducción cercana al 10 % de la plantilla[46].
- **Ryt Bank:** banco malasio que se anuncia como *AI-first*, opera sobre una arquitectura de agentes especializados donde el cliente interactúa en lenguaje natural y los agentes interpretan la intención, orquestan procesos y ejecutan transacciones reales sobre el *core* bancario. El sistema gestiona en torno a 80.000 transacciones al mes, ha reducido procesos que requerían 5-8 pantallas a una sola interacción conversacional, y ha mostrado una mejora significativa en la retención de clientes[47].
- **Walmart:** en sus centros de distribución automatizados, los sistemas de decisión autónoma coordinan almacenamiento, reposición y preparación de pedidos en tiempo real. Walmart reporta que estos centros duplican la capacidad de procesamiento con aproximadamente la mitad de la plantilla frente a centros tradicionales, evidenciando un cambio estructural en la productividad logística[48].
- **Amazon:** su nuevo sistema de gestión logística, Sequoia, orquesta robots, inventario y flujos de pedidos mediante software autónomo que decide dónde almacenar, cuándo mover inventario y cómo abastecer estaciones de recogida. Amazon reporta reducciones de hasta 75 % en el tiempo de manipulación de inventario y hasta 25 % en el tiempo de procesamiento de pedidos en los centros donde está desplegado[49].
- **DHL:** en operaciones de recogida con robots autónomos integrados con el WMS, los sistemas asignan trabajo, optimizan rutas y cierran tareas de forma automática. En despliegues productivos, DHL ha reportado incrementos de productividad de hasta 180 % en unidades por hora, junto con mejoras significativas de calidad y precisión[50].

44 iDanae (2Q25).
45 Deutsche Bank (2025).
46 Financial News London (2025).
47 Ryt Bank (2025).

48 Business Insider (2025a).
49 Amazon (2023).
50 DHL (2024).

¿Pero qué es un sistema agéntico? Cinco capas modulares

Esas capacidades funcionales se materializan en una arquitectura modular concreta (no simplemente un modelo de lenguaje con acceso a herramientas), compuesta por cinco capas interdependientes:

1. **Interfaz y percepción:** recibe objetivos del usuario, entrega resultados finales y percibe eventos del entorno mediante API *gateways*, *endpoints* y conectores de entrada.
2. **Orquestación y planificación**: descompone objetivos complejos en tareas manejables, decide qué agente ejecuta cada tarea y en qué orden, y gestiona colas de prioridades y enrutamiento.
3. **Núcleo de agentes**: trabajadores autónomos, cada uno con rol específico, núcleo cognitivo (LLM) y bucle de control ReAct (*Reason + Act*) que combina razonamiento y acción iterativa.
4. **Herramientas y servicios**: biblioteca de capacidades externas (búsqueda, generación de código, APIs corporativas) con conectividad estandarizada mediante protocolos como MCP y gestión de *prompts*.
5. **Memoria y conocimiento**: almacena información a corto plazo (historial de conversación), a largo plazo (base de datos vectorial con experiencias pasadas) y base de conocimiento corporativo.

Esta arquitectura convierte la autonomía en algo trazable, auditable y gobernable. Cada decisión, cada acción, cada invocación de herramientas queda registrada, permitiendo supervisión humana efectiva y cumplimiento regulatorio.

MCP: el eslabón perdido de la escalabilidad

La arquitectura agéntica enfrenta un problema técnico fundamental: la complejidad exponencial de las integraciones. Tradicionalmente, cada modelo de lenguaje requiere una integración propietaria con cada herramienta. Cambiar de modelo obliga a reescribir todas las integraciones; agregar una nueva herramienta requiere integrarla con todos los modelos existentes. El resultado es deuda técnica exponencial.

Model Context Protocol (MCP) resuelve este problema mediante una capa de abstracción universal. MCP es un estándar abierto que normaliza cómo los modelos de IA interactúan con aplicaciones, fuentes de datos y herramientas externas. Los «servidores MCP» exponen capacidades (recursos, *prompts*, herramientas) que los «clientes MCP» (agentes o modelos) consumen bajo demanda. Una herramienta conectada a MCP es inmediatamente accesible a cualquier agente presente o futuro, sin redesarrollar integraciones.

El impacto de MCP es transformador: pasa de integraciones artesanales a activos reutilizables universales, facilita la transición de prototipos aislados a ecosistemas escalables, permite que los agentes adquieran nuevas capacidades sin redespliegues, y reduce drásticamente el coste de mantenimiento y evolución. Sin MCP o estándares equivalentes, la IA agéntica a escala empresarial no es técnicamente sostenible.

Gobernanza y control: el verdadero reto

Construir agentes es relativamente sencillo con los *framework*s actuales; gobernarlos a escala empresarial es el desafío real. La adopción sostenible requiere equilibrar cuatro capacidades:

- **Tecnología y desarrollo**: orquestación multiagente con coordinación efectiva, gestión de memoria operativa y de largo plazo, arquitecturas modulares que permitan mantenimiento y evolución, e integración segura con sistemas corporativos mediante estándares como MCP.
- **Evaluación continua**: métricas de desempeño, calidad y eficiencia, monitorización de desviaciones y anomalías, control de costes (las llamadas a modelos se multiplican exponencialmente) y trazabilidad completa de acciones y decisiones.
- **Gobernanza y cumplimiento**: supervisión humana mediante mecanismos de aprobación y escalado (teniendo presente que la capacidad humana de supervisión tiene un techo; superado ese límite, la supervisión se vuelve nominal y genera falsa sensación de control), controles y límites explícitos sobre qué acciones puede ejecutar cada agente, transparencia y explicabilidad funcional de las decisiones, y cumplimiento del *AI Act* y políticas internas de riesgo.
- **Industrialización y modelo operativo**: operación 24/7 con mantenimiento continuo, *pipelines* CI/CD para desplegar y actualizar agentes de forma segura, gestión activa de costes en producción, y resiliencia ante fallos o comportamientos inesperados.

Implicaciones estratégicas

La adopción de IA agéntica tiene tres implicaciones estratégicas críticas:

- **Transformación del modelo operativo**: la IA agéntica introduce un cambio estructural donde equipos humanos colaboran con agentes autónomos que operan 24/7 sin supervisión continua. Comprime radicalmente el *time-to-market*: tareas que antes requerían semanas ahora se resuelven en días mediante delegación a agentes especializados.
- **Riesgo de escalada de costes**: a diferencia de la IA generativa tradicional, los agentes multiplican las llamadas a modelos y herramientas. Un prototipo viable puede convertirse en un sistema insostenible económicamente si no se diseña con controles de costes desde el inicio.
- **Industrialización como barrera de entrada**: construir prototipos agénticos es posible con *frameworks* actuales, pero operarlos, escalarlos, mantenerlos, securizarlos y auditarlos en producción requiere capacidades organizativas que la mayoría de empresas aun no han desarrollado. La ventaja competitiva no está en la tecnología, sino en la capacidad de industrializarla con gobernanza efectiva.

IA en robótica y sistemas físicos

De lo digital a la acción en el mundo real

Durante años, la robótica industrial ha operado mediante sistemas programados para tareas repetitivas en entornos altamente controlados: brazos robóticos que ensamblan componentes siguiendo secuencias fijas, vehículos guiados automatizados (AGVs) que siguen rutas predefinidas, sistemas de *pick-and-place* que reconocen objetos en posiciones exactas. Estos robots ejecutan movimientos precisos, pero carecen de adaptabilidad: cualquier cambio en el entorno (un objeto fuera de lugar, una variación en la textura, un obstáculo inesperado) requiere reprogramación o intervención humana.

La integración de IA generativa, los modelos de visión avanzados y el aprendizaje por refuerzo está transformando esta realidad en la industria: solo en 2023 se instalaron más de 276.000 robots industriales en China (Fig. 2), que ya concentra más de la mitad de las instalaciones mundiales, y la proporción de robots colaborativos se ha cuadruplicado en seis años[51]. Los robots actuales perciben su entorno mediante visión por computadora en tiempo real, interpretan instrucciones en lenguaje natural, planifican secuencias de acciones complejas, se adaptan a cambios imprevistos sin reprogramación y aprenden de cada interacción para mejorar continuamente. La IA convierte máquinas industriales rígidas en sistemas autónomos capaces de operar en entornos no estructurados y ejecutar tareas que antes requerían inteligencia humana.

51 Stanford (2025).

Robots humanoides: del laboratorio a la fábrica

La robótica humanoide ha dado un salto cualitativo en los últimos dos años, pasando de demostraciones espectaculares de laboratorio a despliegues industriales reales.

Figure AI, *startup* valorada en aproximadamente 39.000 millones de dólares, completó en 2025 un despliegue de 11 meses de sus robots Figure 02 en la planta de BMW en Spartanburg (Carolina del Sur)[52]. Los dos robots humanoides trabajaron turnos de 10 horas de lunes a viernes, acumulando 1.250 horas de operación, cargando más de 90.000 piezas de chapa metálica con tolerancias de 5 milímetros en 2 segundos por pieza, y contribuyendo a la producción de más de 30.000 vehículos BMW X3.

Figure ha lanzado su tercera generación, Figure 03, diseñada específicamente para producción en volumen. La compañía construyó BotQ, una planta de fabricación dedicada a robots humanoides con capacidad inicial de 12.000 unidades anuales y objetivo de producir 100.000 robots en cuatro años. Figure 03 incorpora carga inalámbrica inductiva (2 kW mediante bobinas en los pies que permiten al robot simplemente pisar una base para recargarse), un sistema de visión rediseñado con doble tasa de refresco y un cuarto del tiempo de latencia, y cámaras integradas en las palmas de las manos para retroalimentación visual redundante durante manipulación fina. La compañía proyecta que estos robots operarán mediante su IA propietaria Helix, entrenada masivamente con datos de teleoperación y demostraciones humanas.

Por su parte, Tesla está acelerando agresivamente la producción de su robot humanoide Optimus. La compañía anunció planes de construir una línea de producción capaz de fabricar un millón de unidades anuales, con inicio esperado hacia finales de 2026[53]. Elon Musk declaró en octubre de 2025 que la versión 3 de Optimus tendrá «manos que son una pieza increíble de ingeniería» con rango completo de movimiento humano (22 grados de libertad) y que el robot será tan realista que «necesitarás tocarlo para creer que es realmente un robot»[54]. Tesla ha producido varios miles de unidades en 2025 para uso interno en sus fábricas (principalmente en tareas de manipulación de baterías y componentes), y planea escalar a 50.000-100.000 unidades en 2026[55]. Musk estima[56] que, en volúmenes superiores al millón de unidades anuales, el coste de producción de Optimus caerá por debajo de 20.000 dólares, aproximadamente la mitad del coste de un *Model Y* a escala equivalente. El precio de venta al público, sin embargo, será significativamente mayor y determinado por demanda de mercado.

Boston Dynamics, referencia histórica en robótica dinámica, retiró en abril de 2024 su Atlas hidráulico (famoso por *backflips* y *parkour*) y lanzó un Atlas completamente eléctrico diseñado para aplicaciones industriales reales[57]. El nuevo Atlas integra actuadores personalizados de alto rendimiento con rango de movimiento que excede capacidades humanas: su cabeza y torso pueden girar 180 grados independientemente, sus articulaciones tienen flexibilidad extrema, y está diseñado para explotar su propia anatomía mecánica, no para limitarse a las posturas humanas, aunque muchas de sus capacidades de control se entrenan a partir de movimiento humano.

52 Figure (2025).
53 Teslarati (2025).
54 Tesla Car World (2025).
55 Fortune (2025b).
56 Notateslaapp (2025).
57 Boston Dynamics (2025a).

Boston Dynamics enfatiza que Atlas priorizará velocidad y eficiencia sobre apariencia antropomórfica.

En agosto de 2025, Boston Dynamics y Toyota Research Institute demostraron[58] Atlas operando mediante *Large Behavior Models* (LBMs): políticas *end-to-end* entrenadas con demostraciones amplias y anotaciones de lenguaje que coordinan locomoción y manipulación simultáneamente. Un solo modelo de comportamiento controla directamente todo el robot, tratando manos y pies de forma casi idéntica, sin separar el control de bajo nivel de locomoción del control de manipulación. En videos públicos[59], Atlas realiza secuencias continuas de tareas de clasificación y embalaje en entornos simulados de fábrica, reaccionando autónomamente a perturbaciones físicas inesperadas (como investigadores cerrando cajas o empujando objetos) sin interrumpir la tarea. Boston Dynamics proyecta pilotos[60] con Hyundai en 2026 y producción comercial limitada a partir de 2027.

Más allá de la manufactura y la logística, la robótica humanoide apunta a un segundo frente de impacto: el cuidado de personas mayores, dependientes o con discapacidad. En economías con envejecimiento estructural acelerado, esta aplicación tiene relevancia estratégica comparable a la industrial, con implicaciones éticas y regulatorias propias que los marcos actuales apenas han comenzado a abordar.

Implicaciones estratégicas y desafíos operativos

La integración de IA en robótica humanoide introduce cambios estructurales en manufactura y logística. La productividad aumenta radicalmente: los robots operan 24/7 sin fatiga, descansos o variación de rendimiento. Los costes operativos recurrentes (energía, mantenimiento preventivo) son predecibles y decrecientes con economías de escala, aunque la inversión inicial sigue siendo significativa.

Sin embargo, surgen desafíos críticos. El impacto en empleo es real y concentrado: tareas manuales repetitivas en manufactura, ensamblaje y manipulación de materiales enfrentan automatización acelerada. Las organizaciones que adoptan robótica humanoide deben gestionar transiciones laborales, programas de *reskilling*, y expectativas regulatorias y sociales crecientes sobre responsabilidad corporativa, desafíos que comparte con la adopción de IA en general, pero que aquí adquieren mayor concentración sectorial y visibilidad social.

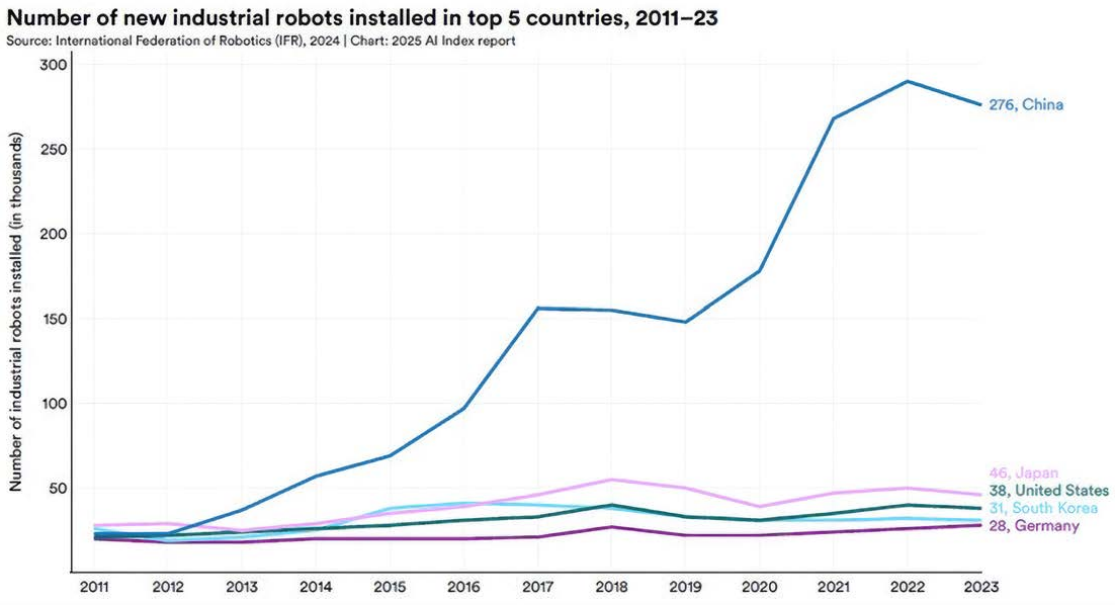
La dependencia de proveedores se intensifica: las empresas que adoptan robots quedan vinculadas a sus ecosistemas propietarios de *hardware*, *software*, actualizaciones de IA y soporte técnico. La obsolescencia tecnológica es rápida: un robot adquirido hoy puede quedar superado en capacidades por la siguiente generación en dos años, lo que plantea interrogantes sobre ciclos de inversión y estrategias de actualización.

Y emergen riesgos operativos nuevos: sistemas autónomos que operan en entornos físicos compartidos con humanos pueden causar lesiones, daños materiales o interrupciones operativas críticas si fallan. La robótica con IA requiere marcos de seguridad robustos (sensores redundantes, sistemas de parada de emergencia, zonas de exclusión dinámicas), protocolos de *fail-safe* certificados, y supervisión humana efectiva incluso en operaciones nominalmente autónomas.

La IA en robótica no es una tendencia futura, es una realidad operativa en fase de industrialización acelerada. Las organizaciones que evalúen estratégicamente cuándo y dónde adoptar robótica humanoide (no todas las tareas justifican la inversión) capturarán ventajas competitivas sostenibles.

58 Boston Dynamics (2025b).
59 Toyota Research Institute (2025).
60 Hyundai (2025).

Fig. 2. Nuevos robots industriales instalados. Fuente: Stanford (2025).

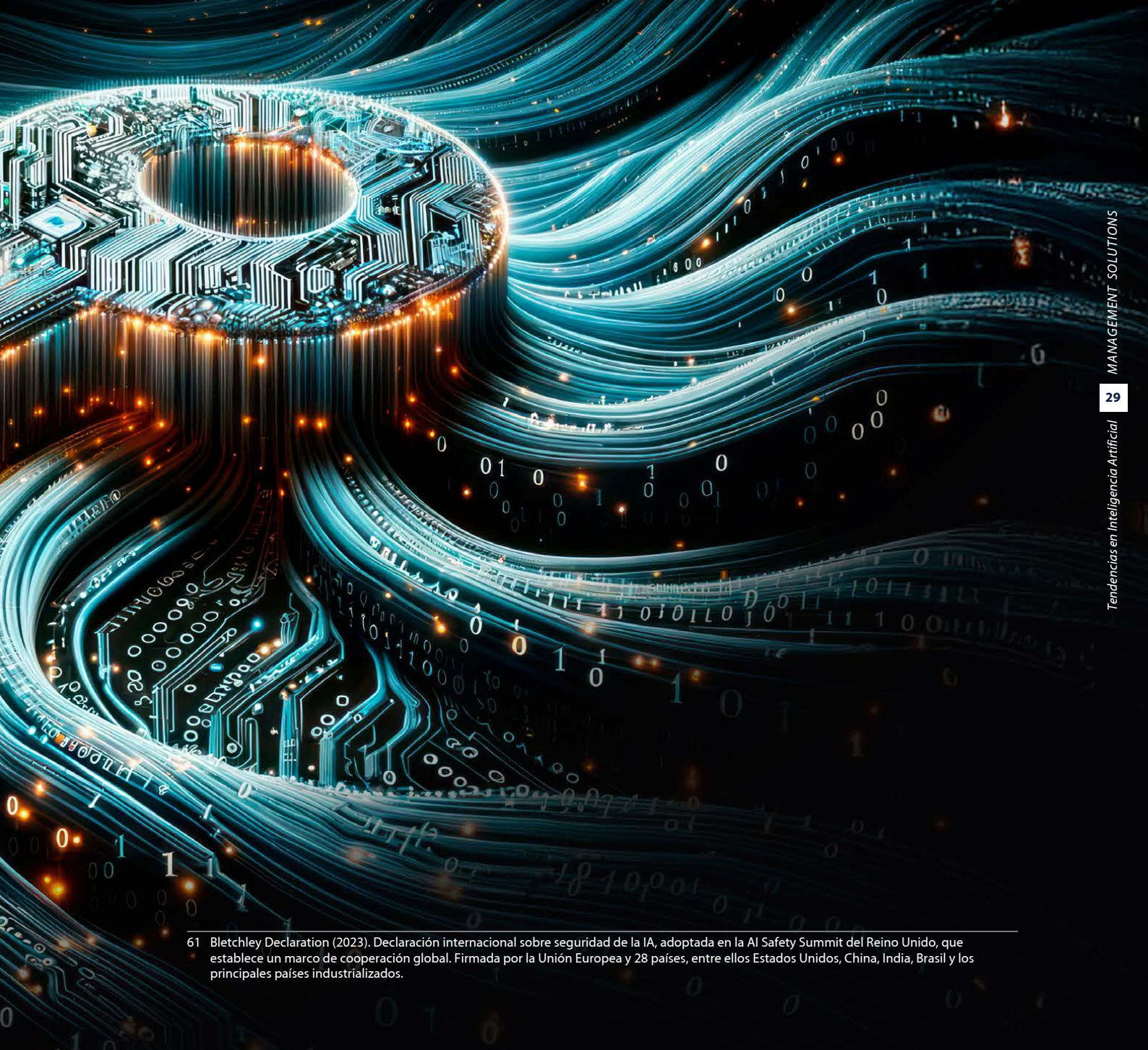


04 | Riesgos, regulación y seguridad de la IA

«Existe la posibilidad de daños graves, incluso catastróficos, ya sean deliberados o no intencionados, derivados de las capacidades más significativas de estos modelos de IA».

Bletchley Declaration[61]

La IA generativa y los sistemas agénticos ofrecen capacidades transformadoras, pero introducen riesgos estructurales que ya no son hipotéticos: se están materializando en producción. La Declaración de Bletchley advirtió explícitamente sobre impactos de la IA potencialmente «catastróficos», y la experiencia acumulada confirma que la lista de riesgos es larga y diversa. Esta sección analiza los riesgos de la IA, los marcos regulatorios y el gobierno de la IA, la batalla entre IA defensiva y adversaria en ciberseguridad, y las tensiones abiertas en privacidad y propiedad intelectual.



61 Bletchley Declaration (2023). Declaración internacional sobre seguridad de la IA, adoptada en la AI Safety Summit del Reino Unido, que establece un marco de cooperación global. Firmada por la Unión Europea y 28 países, entre ellos Estados Unidos, China, India, Brasil y los principales países industrializados.

Riesgos de la IA

La IA no crea el riesgo: lo amplifica

La adopción de la IA no introduce, salvo excepciones, riesgos sustancialmente nuevos – sino que amplifica de forma drástica riesgos ya existentes: operacional, de modelo, tecnológico, de proveedor, legal, reputacional, de cumplimiento, estratégico, social, etc. La diferencia no está en la naturaleza del riesgo, sino en su velocidad de propagación, escala de impacto y dificultad de contención.

Salvo algunas categorías emergentes (como la generación de contenido tóxico, ciertas formas de manipulación cognitiva mediante *deepfakes*, o ataques de *prompt injection*), la mayor parte de los riesgos asociados a la IA son versiones aceleradas, automatizadas y masivas de problemas conocidos. Un sesgo algorítmico es, en esencia, un sesgo humano sistematizado y replicado millones de veces. Una fuga de información por mal uso de un *chatbot* es, a fin de cuentas, una fuga de información.

Cuatro dimensiones interconectadas

En la práctica, estos riesgos se materializan en cuatro grandes dimensiones (Fig. 3):

1. Seguridad y cumplimiento

La IA introduce nuevas superficies de ataque y complica el cumplimiento normativo. La privacidad y seguridad de la información enfrentan amenazas específicas: fugas involuntarias de datos confidenciales, vulnerabilidades técnicas emergentes como *prompt injection* (instrucciones maliciosas embebidas en *inputs* que «engañan» al modelo para ignorar restricciones) o *jailbreaks* (técnicas para eludir controles de seguridad), y exposición accidental de información sensible cuando profesionales utilizan herramientas no securizadas.

La propiedad intelectual se vuelve ambigua: ¿quién es titular del código generado por IA? ¿Qué ocurre cuando un modelo reproduce fragmentos protegidos por derechos de autor? Como ejemplo, el New York Times tiene un litigio abierto con OpenAI/Microsoft por usar contenidos con *copyright* sin licencia, entre otras muchas demandas en curso[62].

La trazabilidad y reproducibilidad se degradan cuando las decisiones críticas dependen de modelos que evolucionan continuamente mediante reentrenamiento automático, haciendo difícil reconstruir exactamente qué versión del sistema produjo qué resultado en qué momento.

Y el incumplimiento normativo tiene ahora consecuencias financieras directas y visibles: el *AI Act* europeo[63] establece multas de hasta 35 millones de euros o el 7 % de la facturación global anual, lo que convierte el cumplimiento en IA en riesgo material de primer orden, superior incluso al de protección de datos[64].

2. Calidad, fiabilidad, *lock-in* y costes

Como ya se ha dicho, la IA generativa produce por diseño salidas plausibles, no necesariamente correctas. Las alucinaciones (generación de información falsa presentada con confianza) no son *bugs* ocasionales, sino comportamientos intrínsecos del diseño actual de los modelos. El comportamiento no determinista implica que el mismo *input* puede producir *outputs* distintos en el mismo modelo, lo que complica la validación, auditoría y certificación de procesos críticos.

El *model drift* silencioso representa uno de los riesgos más insidiosos: un modelo que funcionaba correctamente puede degradarse en poco tiempo porque la distribución de datos cambió, sin que existan mecanismos de monitorización que generen alertas tempranas. Un modelo puede seguir operando con apariencia de normalidad hasta que alguien detecta manualmente las anomalías en los resultados.

La dependencia excesiva de la IA en tareas críticas crea fragilidad operativa: si el sistema de IA falla, cae o se vuelve prohibitivamente caro, ¿puede la organización seguir operando? ¿Existen procedimientos de contingencia humanos? Y la pérdida de control humano efectivo ocurre cuando las decisiones se delegan a sistemas cuyo razonamiento interno es opaco incluso para quienes los desarrollan y operan.

Más aún, las organizaciones construyen dependencias estructurales sobre pocos proveedores de modelos (OpenAI, Anthropic, Google) e infraestructura (AWS, Azure, GCP). Cambios en *pricing*, términos de servicio o caídas operativas pueden paralizar procesos críticos simultáneamente. Migrar entre ecosistemas implica reescribir integraciones, reentrenar *workflows*, recertificar *compliance* y asumir costes prohibitivos, lo que no siempre es factible; la diversificación de proveedores, aunque costosa, es la respuesta estructural a este riesgo.

A los riesgos anteriores se añade una dimensión económica que los marcos de gestión tradicionales infravaloran. Los costes unitarios de inferencia caen estructuralmente (aproximadamente 10x cada 12 meses) pero esa caída no protege contra el crecimiento exponencial del volumen: los sistemas agénticos tienen estructuras de coste con escalabilidad no lineal, cada agente multiplica llamadas a modelos y herramientas, y sin monitorización de *tokens* en tiempo real ni límites de gasto explícitos desde el diseño, un prototipo viable puede convertirse en un sistema insostenible antes de que nadie lo detecte[65]. A esto se suma la incertidumbre sobre el retorno de la inversión: la pregunta de si el gasto masivo en IA generará el valor esperado no tiene aún respuesta consolidada, y muchas organizaciones avanzan impulsadas por la presión competitiva más que por un caso de negocio robusto. Gartner sintetiza ambos riesgos en una predicción: más del 40 % de los proyectos agénticos serán cancelados antes de 2027 por escalada de costes, valor de negocio poco claro y controles de riesgo inadecuados[66]. El *lock-in* agrava el problema estructuralmente: la dependencia de un único proveedor elimina la capacidad de migrar a arquitecturas más eficientes cuando estén disponibles.

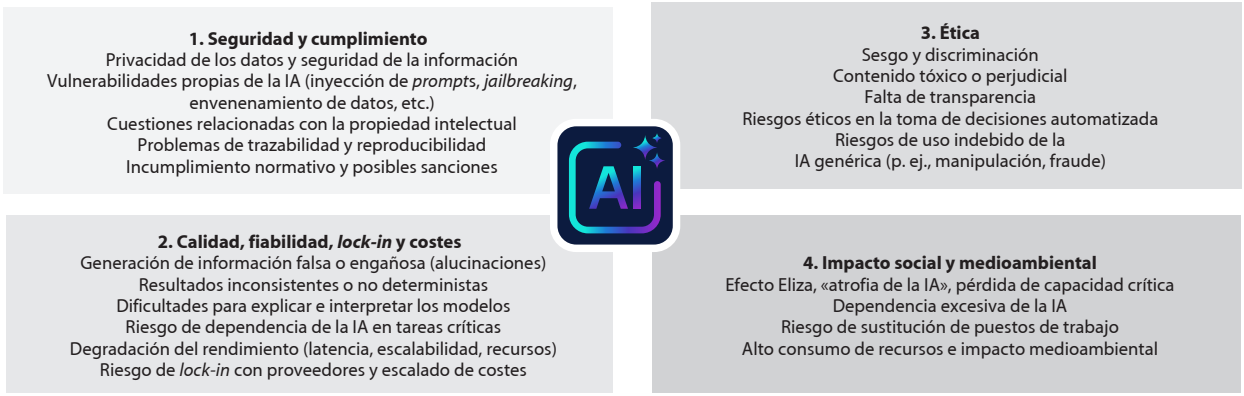
3. Ética y toma de decisiones automatizada

Los sesgos algorítmicos amplifican sesgos presentes en datos históricos: si un modelo de selección de personal se entrena con

62 Brown (2025).
63 AI Act (2024).
64 Management Solutions (4Q24).

65 La estimación de costes antes del desarrollo y la monitorización continua en producción son dos capacidades que las organizaciones están comenzando a formalizar en sus marcos de gobierno de IA. GenMS Tracker, solución propietaria de Management Solutions, aborda ambas dimensiones: permite estimar el coste de desarrollo y operación de un sistema de IA a partir de una descripción en lenguaje natural, y monitoriza en tiempo real el consumo en producción con alertas ante desviaciones respecto al presupuesto definido.
66 Gartner (2025b).

Fig. 3. Algunos riesgos de la IA generativa.



datos de una organización históricamente sesgada hacia ciertos perfiles, el modelo sistematizará y escalará ese sesgo[67]. La falta de transparencia y las dificultades de explicabilidad complican el cumplimiento de requisitos regulatorios que exigen que las decisiones automatizadas sean comprensibles y justificables, especialmente cuando afectan derechos fundamentales.

Las decisiones automatizadas con impacto humano directo (aprobación de créditos, diagnósticos médicos, evaluación de riesgos penales, selección de candidatos) plantean problemas de responsabilidad: ¿quién responde cuando el modelo comete un error con consecuencias graves? ¿El proveedor del modelo base, el equipo que lo personalizó, el usuario que escribió el *prompt*, el comité que aprobó su despliegue? La cadena de responsabilidad se vuelve compleja, lo que genera vacíos de *accountability* que ni la regulación ni la práctica han resuelto completamente.

4. Impacto social y medioambiental

La transformación del empleo ya no es especulativa: tareas manuales y cognitivas rutinarias enfrentan automatización acelerada, y las organizaciones deben gestionar transiciones laborales, programas de *reskilling* y expectativas regulatorias y sociales crecientes sobre responsabilidad corporativa.

La dependencia cognitiva y erosión de capacidades críticas representa un riesgo de primer orden: si una generación entera de profesionales aprende a trabajar exclusivamente con IA como intermediario, ¿mantienen la intuición y el conocimiento suficientes para valorar críticamente los resultados? ¿Qué ocurre cuando la IA no está disponible? Y la presión sobre recursos y huella medioambiental no es trivial: entrenar modelos avanzados consume energía equivalente a miles de hogares durante meses, y el escalado masivo de inferencia plantea interrogantes sobre sostenibilidad a largo plazo[68].

Amplificación no lineal

La IA introduce un fenómeno clave en la gestión del riesgo: la amplificación no lineal. Un fallo menor (un sesgo en datos, un *prompt* mal diseñado, una mala configuración de permisos) puede escalar en minutos y afectar simultáneamente a procesos, clientes, reguladores y reputación. Cuanto mayor es la autonomía y la integración de la IA en procesos críticos, mayor es el radio de impacto del fallo, lo que exige diseñar desde el inicio mecanismos de contención proporcionales a esa escala.

Un ejemplo concreto: un modelo de atención al cliente que, tras un cambio menor en el *prompt* del sistema, empieza a revelar información confidencial de otros clientes en «tan solo» el 0,01 % de las conversaciones, lo que es casi indetectable en las pruebas. En un sistema que gestiona 100.000 interacciones diarias, esto significa 10 filtraciones al día antes de que se detecte el patrón. Cuando se identifica el problema, ya se han producido cientos de incidentes, cada uno con implicaciones regulatorias, contractuales y reputacionales.

IA en la taxonomía de riesgos corporativos

Las organizaciones están adoptando uno de tres enfoques para ubicar el riesgo de IA en su marco de riesgos corporativo:

- Tratarlo como riesgo de primer nivel, creando una categoría nueva en la taxonomía (enfoque muy poco frecuente, pero útil para ganar visibilidad ejecutiva en fases tempranas de adopción masiva).
- Tratarlo como riesgo de segundo nivel, vinculado a riesgo tecnológico, de modelo, u operacional, dentro de la taxonomía existente.
- No tratarlo como categoría independiente, sino como *driver* transversal que amplifica riesgos existentes (riesgo de modelo, de proveedor, reputacional, etc., amplificado por IA).

En la práctica, la etiqueta es menos importante que la capacidad de la organización para identificar, prevenir, controlar y mitigar estos riesgos de forma sistemática y continua, con métricas claras, responsabilidades asignadas y mecanismos de escalado definidos.

Implicación estratégica

La gestión del riesgo en IA se ha convertido en una decisión estratégica que condiciona la velocidad de adopción, la viabilidad operativa y la credibilidad institucional de la organización.

Las entidades que enfocan la IA principalmente como proyecto de innovación tecnológica encuentran fricciones con el marco regulatorio, incidentes operativos y tensiones reputacionales. Integrar la IA desde su concepción dentro de la arquitectura de gobierno, control interno y gestión de riesgos permite escalar con mayor estabilidad, menor fricción y mayor confianza por parte del regulador, del mercado y de los propios profesionales.

La ventaja competitiva sostenible surge de una mayor capacidad para gobernar la IA: marcos de control bien definidos, responsabilidades explícitas, monitorización continua y cultura organizativa que preserve el juicio humano como elemento central de las decisiones críticas.

67 Management Solutions (2023).
68 iDanae (1Q24).

Regulación, supervisión y estándares de IA

La IA, una actividad regulada

La aceleración de la adopción de la IA y la constatación de sus riesgos han desencadenado una respuesta regulatoria global sin precedentes. A diferencia de ciclos tecnológicos anteriores, donde la regulación reaccionaba con años de retraso, la IA está siendo regulada en paralelo a su despliegue masivo, precisamente porque sus riesgos sistémicos ya se están materializando.

Europa ha asumido el liderazgo normativo con el *AI Act*[69], el primer marco legal integral sobre IA. Le siguen iniciativas relevantes en Estados Unidos, China, Reino Unido, Canadá y otros países, junto con un ecosistema creciente de estándares técnicos y marcos voluntarios que buscan operacionalizar principios de gobernanza, seguridad y ética[70].

La consecuencia para las organizaciones es clara: la gestión de IA deja de ser una cuestión tecnológica para convertirse en un dominio regulatorio estructural, comparable en impacto al de protección de datos, mercados financieros o seguridad operacional.

El AI Act: la regulación más prescriptiva

El Reglamento Europeo de IA (Reglamento (UE) 2024/1689) establece un modelo de regulación basado en el riesgo, estructurando las obligaciones según el impacto potencial de los sistemas sobre los derechos fundamentales, la seguridad y el orden público.

El marco clasifica los sistemas de IA en cuatro categorías principales:

1. Riesgo inaceptable: sistemas prohibidos (manipulación cognitiva, puntuación social, ciertas formas de vigilancia biométrica).
2. Riesgo alto: sistemas utilizados en ámbitos críticos como crédito, empleo, educación, infraestructuras, justicia, sanidad, biometría, entre otros. Es el núcleo del *AI Act*.
3. Riesgo limitado: sistemas que requieren obligaciones de transparencia.
4. Riesgo mínimo: sistemas de uso libre bajo algunos principios generales.

Para los sistemas de alto riesgo, el *AI Act* introduce un conjunto de obligaciones estructurales:

- Sistema de gestión de riesgos documentado.
- Gobierno y control de calidad de datos.
- Documentación técnica exhaustiva.
- Registro de *logs* y trazabilidad.
- Supervisión humana efectiva.
- Requisitos de precisión, robustez y ciberseguridad.
- Evaluación de conformidad previa al despliegue y vigilancia continua post-mercado.

Las sanciones por incumplimiento grave del *AI Act* alcanzan los 35 millones de euros o el 7 % de la facturación global anual, superando incluso el régimen del GDPR (que alcanza el 4 %).

Supervisión en Europa: de la innovación al control permanente

El *AI Act* crea una arquitectura institucional nueva:

- *AI Office*: órgano técnico central de la Comisión para IA, sobre todo GPAI.
- Autoridades nacionales de supervisión: supervisan y sancionan el cumplimiento del *AI Act* en cada país.
- *AI Board*: foro de coordinación e interpretación común entre la Comisión y las autoridades nacionales.
- Mecanismos de cooperación transfronteriza: reglas para coordinar casos e investigaciones que afecten a varios Estados miembros.

La supervisión no se limitará a auditorías puntuales: se establece un modelo de vigilancia continua, con obligaciones de *reporting*, gestión de incidentes graves, retirada de sistemas no seguros y poderes de inspección comparables a los de reguladores financieros.

En la práctica, la arquitectura supervisora del *AI Act* aún está en fase de construcción. El *AI Board*, órgano central de coordinación entre Estados miembros y la Comisión, celebró su primera reunión formal[71] en septiembre de 2024, y se ha centrado en cuestiones organizativas, códigos de práctica para modelos GPAI, y coordinación de autoridades nacionales. Las actas revelan un proceso aún en fase de organización: selección de presidencia, creación de subgrupos y discusión sobre entregables prioritarios.

Por su parte, las autoridades nacionales de supervisión apenas han sido designadas. España creó la AESIA (Agencia Española de Supervisión de la Inteligencia Artificial)[72] en agosto de 2023, convirtiéndose en la primera autoridad de IA de Europa, y comenzó

69 AI Act (2024).
70 Management Solutions (4Q24a).
71 AI Board (2026).
72 AESIA (2026).

operaciones efectivas en febrero de 2025 con funciones de supervisión de sistemas prohibidos. Otros Estados miembros están en fases más tempranas de designación de autoridades.

El ecosistema de estándares: de los principios a la operación

En paralelo a la regulación, se consolida un entramado de estándares técnicos y de gestión que buscan establecer normas para la práctica operativa de la IA; entre ellos:

- ISO/IEC 42001: establece requisitos para un sistema de gestión de IA (AIMS), al estilo ISO 27001/9001, para gobernanza y uso responsable de la IA.
- ISO/IEC 23894: proporciona un marco detallado de gestión de riesgos de IA a lo largo del ciclo de vida, pensado para integrarse con *frameworks* de gestión de riesgos corporativos.
- ISO/IEC 5259: marco y métricas para calidad del dato en IA.
- NIST *AI Risk Management Framework*: marco centrado en la gestión de riesgos de IA, organizado en las funciones *Govern-Map-Measure-Manage* y orientado a características de *trustworthy A*I.
- OECD *AI Principles*: principios de alto nivel (valores humanos, transparencia, robustez, *accountability*, inclusión) que han influido directamente en el *AI Act* y otros marcos regulatorios.

Estos estándares cumplen una función esencial: establecen controles concretos, procesos auditables y métricas verificables. Para muchas organizaciones, están constituyendo la base técnica de sus programas de cumplimiento.

Fragmentación geopolítica y complejidad operativa

El enfoque europeo basado en riesgo y vinculante no se está replicando globalmente:

- **Estados Unidos** carece de marco federal; dispone de un enfoque sectorial fragmentado por agencias (FTC, FDA, EEOC), sin equivalente al *AI Act*, complementado por órdenes ejecutivas y guías, y de regulación en varios de los Estados (p. ej., California, Colorado, Connecticut, Utah) [73].
- **China** articula la IA dentro de una estrategia estatal de «soberanía digital», con normas específicas sobre algoritmos, *deepfakes* y modelos generativos, fuerte control de datos y licencias obligatorias para ciertos sistemas de alto impacto[74].
- **Reino Unido** mantiene un enfoque pro-innovación sin una ley horizontal de IA, apoyándose en autoridades sectoriales y principios comunes de regulación de IA, con guías y *sandboxes* regulatorios[75].
- **Brasil** aprobó en el Senado (diciembre 2024) un proyecto de ley estructuralmente similar al *AI Act*: enfoque basado en riesgo (prohibido/alto/limitado/mínimo), obligaciones específicas para sistemas de alto riesgo, multas hasta R$50 millones o 2 % de facturación. El proyecto está pendiente de aprobación en la Cámara de Diputados, con entrada en vigor prevista un año después de promulgación[76].
- **México** carece de regulación: desde 2020 se han presentado más de 60 proyectos de ley sin aprobación. En febrero de 2025 se introdujo una reforma constitucional para otorgar competencia federal en IA, que debe traducirse en Ley General. Hasta entonces, no existe marco legal específico, solo guías voluntarias alineadas con UNESCO y OCDE[77].
- **Australia** mantiene un enfoque voluntario y basado en principios: *AI Ethics Principles* (2019, 8 principios voluntarios) y *Guidance for AI Adoption* (octubre 2025, sustituyendo el *Voluntary AI Safety Standard* de 2024). No hay legislación vinculante específica para IA, y el gobierno ha preferido reforzar leyes existentes (privacidad, consumidor, sectorial)[78].

Implicaciones estratégicas

Los estudios coinciden[79] en que la divergencia entre estos modelos (UE más prescriptiva, EEUU fragmentado y sectorial, China altamente centralizada, Reino Unido y Australia más *principle-based*) obliga a las compañías globales a segmentar productos, modelos y procesos de cumplimiento por jurisdicción.

Esto se traduce[80] en arquitecturas de IA multinivel (gobernanza, datos, MLOps, documentación) diseñadas para mapear y reconciliar simultáneamente requisitos divergentes, lo que incrementa la complejidad operativa de forma sin precedentes para muchos sectores tradicionalmente menos regulados.

73 Patel (2025).
74 Cambridge (2025).
75 UK Government (2023).
76 Inter-Parliamentary Union (2025).
77 Covington (2025).
78 Australian Government (2025).
79 Oxford (2025).
80 World Bank (2024).

IA y ciberseguridad

Una batalla con nuevas reglas y nuevas superficies de ataque

La IA está transformando radicalmente la ciberseguridad en tres dimensiones simultáneas: amplifica las capacidades ofensivas de los atacantes, potencia las defensas de las organizaciones y, al mismo tiempo introduce nuevas vulnerabilidades que requieren protección específica.

IA ofensiva: automatización y adaptación a escala industrial

Los atacantes han adoptado IA con velocidad sorprendente, transformando métodos tradicionales en amenazas cualitativamente diferentes. El impacto es cuantificable[81]: en 2025 se registraron más de 28 millones de ciberataques potenciados por IA globalmente, un incremento del 47 % interanual; el sector financiero fue el más afectado, con un 33 % de estos ataques; y el 87 % de las organizaciones experimentaron al menos un ataque asistido por IA en los últimos 12 meses[82].

***Phishing* hiperpersonalizado a escala masiva.** Los ataques de *phishing* generados por IA aumentaron un 1.265 % en un año desde el lanzamiento de ChatGPT[83], y más del 80 % de los *emails* de *phishing* utilizan ahora modelos de lenguaje para generación de texto[84]. La diferencia cualitativa es dramática: mientras el *phishing* tradicional alcanza tasas de éxito del 12 %, las campañas generadas por IA logran 54 % de clics[85]. Los LLMs permiten crear mensajes personalizados analizando perfiles públicos, estilo de escritura corporativa y contextos específicos de la víctima, a velocidades y volúmenes imposibles manualmente.

***Malware* polimórfico adaptativo.** El 76 % del *malware* detectado en 2025 exhibe características polimórficas potenciadas por IA[86]. A diferencia del *malware* polimórfico tradicional (que muta mediante ofuscación o encriptación rutinaria), el *malware* generado por IA reescribe dinámicamente su código en tiempo real, manteniendo funcionalidad idéntica, pero con firmas completamente diferentes. Algunas variantes avanzadas generan versiones únicas cada 15 segundos durante un ataque. Esto derrota a los sistemas de detección basados en firmas estáticas, que históricamente han sido la base de los antivirus tradicionales.

Más preocupante: la IA no solo genera variantes, sino que adapta el comportamiento. Los modelos de *machine learning* embebidos en *malware* analizan el entorno de ejecución, detectan sistemas de monitorización y ajustan sus tácticas en tiempo real para evadir la detección.

***Deepfakes* y manipulación de identidad.** Según análisis de ciberseguridad[87] basados en el IC3 2025, los ataques de Business Email Compromise (BEC) asistidos por IA habrían aumentado alrededor de un 37 %, combinando texto, audio y vídeo sintéticos para suplantar a directivos. El caso más notorio: un *deepfake* de audio del Ministro de Defensa italiano que causó pérdidas financieras significativas[88]. En una encuesta, el 85 % de las organizaciones reportaron haber experimentado algún ataque mediante *deepfake* en 2025[89].

***Dark LLMs* y herramientas ofensivas especializadas.** Han proliferado modelos de lenguaje modificados específicamente para cibercrimen: HackerGPT, WormGPT, GhostGPT, FraudGPT. Estos sistemas, creados mediante *jailbreaking* de modelos éticos o modificación de modelos *open-source*, se comercializan en foros de *dark web* con modelos de suscripción y soporte técnico. Generan *scripts* maliciosos, *exploits* y campañas de ingeniería social sin restricciones éticas.

IA defensiva: detección comportamental y respuesta automatizada

Las organizaciones están respondiendo con IA defensiva igualmente sofisticada. El 51 % de las empresas utilizan ahora IA o automatización en seguridad, y la adopción acelera rápidamente ante la evidencia de ROI demostrable[90].

Análisis comportamental y detección de anomalías. Los sistemas de *User and Entity Behavior Analytics* (UEBA) potenciados por IA establecen líneas base dinámicas de comportamiento normal para usuarios, dispositivos y aplicaciones, analizando miles de millones de eventos diarios. En lugar de buscar firmas conocidas, detectan desviaciones sutiles de patrones establecidos. Esta capacidad resulta crítica contra *malware* polimórfico y ataques *zero-day*: en entornos de alto riesgo, los sistemas basados en IA alcanzan tasas de detección de hasta el 98 %[91] frente a amenazas conocidas o que exhiben patrones anómalos reconocibles. Frente a ataques genuinamente inéditos (sin firma ni patrón comportamental previo) la detección basada en comportamiento reduce el riesgo pero no elimina la incertidumbre: la IA defensiva no reconoce la amenaza nueva, sino que detecta su desviación de la norma, lo que implica que ataques suficientemente cautelosos o bien diseñados para mimetizar comportamiento legítimo pueden eludir esa detección inicial.

Plataformas SIEM/XDR/SOAR con IA integrada. Las plataformas actuales de *Security Information and Event Management* (SIEM), *Extended Detection and Response* (XDR) y *Security Orchestration, Automation and Response* (SOAR) integran IA nativamente para correlacionar eventos entre sistemas dispares, reducir falsos positivos (hasta 95 % de reducción en implementaciones maduras), y automatizar la respuesta. CrowdStrike reporta[92] que su plataforma Falcon analiza 4,7 billones de eventos diarios con threat hunting 24/7 potenciado por IA. Microsoft Sentinel ha demostrado[93] reducciones del 30 % en tiempo medio de respuesta (MTTR) mediante correlación y análisis comportamental basados en IA.

Impacto económico demostrable. Según IBM[94], las organizaciones que utilizan IA y automatización extensivamente en seguridad reducen los costes medios de las brechas en $1,9 millones (más de 50 % menos que organizaciones sin IA) y acortan los ciclos de contención en 80 días de media. Las organizaciones con

81 SQ Magazine (2025).
82 SoSaft (2025).
83 PR Newswire (2023).
84 KnowBe4 (2025).
85 AICerts (2026).
86 WatchGuard (2025).
87 The Network Installers (2025).
88 Phishcare (2025).
89 Ironscales (2025).
90 IBM (2025).
91 Proofpoint (2025).
92 CrowdStrike (2025).
93 Microsoft (2026).
94 IBM (2025).

plataformas *AI-driven* detectan amenazas un 60 % más rápido[95] y alcanzan un 95 % de precisión en la detección.

Multiplicación de capacidades. La IA actúa como multiplicador de capacidades para equipos de seguridad: automatiza el triage de alertas (las organizaciones enfrentan un promedio de 4.500 alertas diarias[96]), ejecuta *playbooks* de respuesta automática para amenazas conocidas y permite que analistas junior operen a niveles de efectividad superiores. El 95 % de profesionales de seguridad reporta que la IA mejora su velocidad y eficiencia para prevenir, detectar, responder y recuperarse de ataques[97].

La asimetría y el dilema estratégico

A pesar de capacidades defensivas avanzadas, persiste una asimetría preocupante: la mayoría de las compañías carece actualmente de madurez suficiente para contrarrestar amenazas avanzadas potenciadas por IA. El 78 % de CISOs afirman que las amenazas impulsadas por IA tienen ahora «impacto significativo» en sus organizaciones[98].

La ciberseguridad se ha convertido en una batalla de IA contra IA, donde ambos lados operan a velocidades de máquina. Los atacantes automatizan el reconocimiento, generan *exploits* personalizados y adaptan sus tácticas en tiempo real. Los defensores correlacionan *terabytes* de telemetría, predicen vectores de ataque y ejecutan contención autónoma. El diferencial competitivo ya no reside en tener IA, sino en la sofisticación de los modelos, la calidad de los datos de entrenamiento, la velocidad de actualización de inteligencia de amenazas y la capacidad de integración entre superficies de ataque.

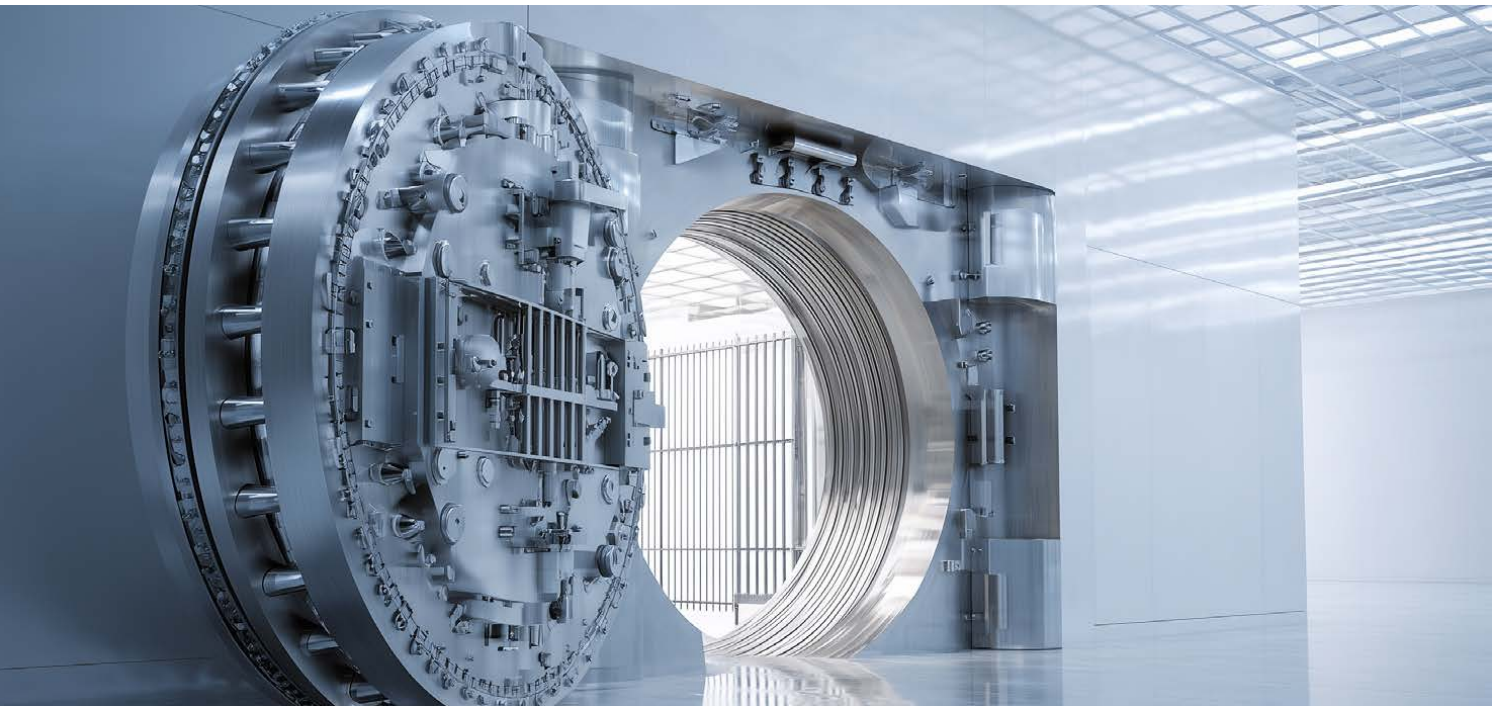
Securización de la IA: vulnerabilidades propias

Más allá de la batalla entre IA ofensiva y defensiva, emerge una tercera dimensión crítica: la securización de los propios sistemas de IA, que introducen vectores de ataque sin precedentes en el *software* tradicional. Los modelos de IA son vulnerables a ataques adversarios específicos: envenenamiento de datos de entrenamiento (*data poisoning*), donde atacantes inyectan datos maliciosos para degradar el modelo; evasión adversaria mediante perturbaciones imperceptibles que engañan al modelo durante la inferencia; y extracción de modelos vía consultas repetidas para robar propiedad intelectual[99].

Los LLMs añaden vectores adicionales documentados por OWASP[100]: *prompt injection* (instrucciones maliciosas embebidas en *inputs* que alteran el comportamiento del modelo), *insecure output handling* (aplicaciones que confían ciegamente en *outputs* sin validación), *training data poisoning, model denial of service* y vulnerabilidades de *supply chain*. NIST publicó en 2024 guías específicas para desarrollo seguro de IA generativa[101], extendiendo su SSDF con controles diferenciados para cada fase del ciclo de vida. La securización efectiva requiere curación rigurosa de *datasets*, adversarial *robustness testing*, validación de *inputs/outputs*, *sandboxing* y *red teaming* específico de ML. Los equipos de seguridad deben incorporar *expertise* en ataques adversarios; los marcos de gobierno deben contemplar la IA como superficie de ataque independiente con controles propios[102].

El cibercrimen global alcanza billones anuales. Las organizaciones deben mantener una gobernanza robusta: los propios sistemas de IA defensivos son ahora objetivo de ataques (envenenamiento de modelos, *prompt injection*, evasión adversaria), creando una capa meta de riesgo que requiere protección especializada.

95 Jumpcloud (2025).
96 HelpNetSecurity (2025).
97 Darktrace (2025).
98 Darktrace (2025).
99 NIST (2024a).
100 OWASP (2025).
101 NIST (2024a).
102 MITRE (2025).



IA, privacidad y propiedad intelectual

Un conflicto estructural

La adopción masiva de IA generativa reabre debates fundamentales sobre privacidad y propiedad intelectual, enfrentando los modelos de negocio de sistemas que requieren datos masivos con marcos legales diseñados para minimización y control individual. Las tensiones no son meramente técnicas: reflejan choques estructurales entre la lógica operativa de LLMs y los principios que rigen protección de datos y derechos de autor.

Privacidad en LLMs: riesgos sistémicos en todo el ciclo de vida

En abril de 2025, el European Data Protection Board (EDPB) publicó[103] un informe exhaustivo sobre riesgos de privacidad en LLMs, desarrollado bajo su programa Support Pool of Experts. El documento identifica que cada fase del ciclo de vida de un LLM introduce riesgos específicos de privacidad y protección de datos:

- **Memorización involuntaria y fuga de datos personales.** Los LLMs pueden memorizar fragmentos de datos personales presentes en *datasets* de entrenamiento y reproducirlos posteriormente en *outputs* generados. Este fenómeno no es un *bug* ocasional, sino un comportamiento intrínseco: el modelo almacena patrones estadísticos que incluyen datos sensibles. El EDPB documenta casos donde *prompts* específicos han logrado extraer información personal (nombres, *emails*, números de teléfono, datos médicos) que estaban en los datos de entrenamiento. La magnitud del problema escala con el tamaño del modelo y la sensibilidad de los datos procesados.
- **Re-identificación y perfilado no intencional.** Aunque los datos se anonimicen antes del entrenamiento, las técnicas de inferencia pueden re-identificar individuos combinando múltiples *outputs* del modelo. El EDPB advierte que los LLMs pueden generar perfiles detallados de individuos sin procesamiento explícito de datos personales identificables, lo que viola los principios GDPR de minimización y finalidad.
- ***Feedback loops* sin salvaguardas.** Las interacciones de usuarios con *chatbots* se almacenan frecuentemente para el *fine-tuning* posterior de los modelos, de modo que los datos sensibles revelados en conversaciones se incorporan al modelo sin consentimiento explícito ni garantías de eliminación posterior.
- **Incompatibilidad estructural con GDPR.** El informe de EDPB subraya tensiones irresolubles con los principios GDPR:
  - **Minimización de datos:** los LLMs requieren *datasets* masivos, lo que contradice frontalmente la minimización.
  - **Derecho al olvido:** no existen aún métodos robustos para «des-entrenar» modelos selectivamente (las técnicas emergentes como *machine unlearning* están en fase experimental).
  - **Transparencia:** las arquitecturas *transformer* son *black boxes* donde rastrear el origen de *outputs* específicos es técnicamente complejo.
  - **Consentimiento:** los datos scrapeados de internet raramente incluyen consentimiento para su uso en el entrenamiento de IA.

103 EDPB (2025).

- **Evaluación de impacto obligatoria.** El EDPB concluye que, dada la naturaleza sistémica del procesamiento en LLMs, realizar *Data Protection Impact Assessment* (DPIA) según GDPR Art. 35 no es solo recomendable sino obligatorio en la mayoría de los casos, especialmente cuando los LLMs procesan datos sensibles o toman decisiones con impacto sobre individuos.
- **Mitigaciones técnicas con coste.** El reporte propone medidas como *differential privacy* (añadir ruido estadístico para impedir la identificación), *federated learning* (entrenar modelos sin centralizar datos), *Retrieval-Augmented* Generation (RAG, que separa el conocimiento actualizable de la memoria estática del LLM), y *retrospective logging minimization* (minimizar los datos en *logs* de sistema). Sin embargo, todas estas técnicas implican una reducción de la precisión, un coste computacional elevado o una funcionalidad reducida.

Propiedad intelectual: litigios masivos y colapso de infraestructura

La World Intellectual Property Organization (WIPO) ha dedicado múltiples sesiones de su «*WIPO Conversation on IP and AI*»[104] a analizar el impacto de la IA generativa sobre el copyright. Las últimas sesiones se centraron específicamente en infraestructura para gestión de derechos, atribución y compensación en la era de la IA generativa.

Datos de entrenamiento: ¿uso legítimo o infracción masiva? El debate central no está resuelto. Las compañías de IA argumentan que entrenar modelos con contenido protegido constituye *fair use* (uso legítimo sin licencia), análogo a cómo los humanos aprenden leyendo. Los titulares de derechos argumentan que se trata de una copia no autorizada a escala industrial con intención comercial que crea sustitutos de mercado para el contenido original.

Los litigios se multiplican; por citar algunos ejemplos:

- **New York Times vs. OpenAI/Microsoft**: el NYT demanda por el uso de millones de artículos sin consentimiento, argumentando que los modelos crean sustitutos de mercado que desvían tráfico de su *paywall* y que generan alucinaciones que dañan su reputación. Reclama miles de millones de dólares en daños.
- **Getty Images vs. Stability AI**: Getty acusa del uso sin consentimiento de más de 12 millones de fotografías para entrenar Stable Diffusion, incluyendo una infracción de la marca registrada (dado que el modelo replica la marca de agua de Getty)[105].
- **Artistas vs. Midjourney/Stability AI**: los artistas argumentan que sus obras fueron *scrapeadas* sin permiso para entrenar modelos generadores de imágenes.
- **Discográficas vs. Anthropic**: Universal Music Group (UMG) y otras discográficas demandan por infracción masiva del uso de letras de canciones.

A diciembre de 2025, hay en curso[106] más de 72 litigios de copyright activos contra compañías de IA. Hasta ahora, tres jueces han emitido decisiones preliminares sobre el uso legítimo: dos sentencias favorables a las compañías de IA[107], una contraria[108]. No se esperan decisiones definitivas hasta el verano de 2026 como mínimo.

Titularidad de los *outputs*: un vacío legal. ¿Quién es el titular del contenido generado por IA? La mayoría de las jurisdicciones (incluyendo la US *Copyright Office*) mantiene que el copyright requiere autoría humana con «*input* creativo suficiente». El contenido generado puramente por IA sin intervención humana significativa cae en el dominio público. Pero los límites son difusos: ¿cuánta intervención humana (*prompt engineering*, selección, *post*-edición) es «suficiente»?

Colapso de la infraestructura de *copyright*. La WIPO advierte de que los sistemas de gestión colectiva de derechos, diseñados para volúmenes manejables de obras humanas, colapsan ante los billones de *outputs* generados por IA diariamente. No existen infraestructuras escalables para el seguimiento, la atribución y la compensación. Las últimas sesiones de la WIPO exploraron la necesidad de nuevas infraestructuras técnicas y marcos regulatorios, pero las soluciones prácticas permanecen especulativas.

Fragmentación y ausencia de consenso global

La fragmentación regulatoria amplifica la incertidumbre. La UE exige transparencia sobre los datos de entrenamiento (*AI Act Art.* 53)[109], Estados Unidos carece de legislación federal específica, China impone un control estatal estricto sobre datos y algoritmos. Las compañías que operan globalmente enfrentan requisitos no armonizados.

Las tecnologías de preservación de la privacidad (*differential privacy, federated learning, homomorphic encryption*) ofrecen rutas técnicas para reconciliar la privacidad con la utilidad de la IA, pero están lejos de la adopción masiva: son costosas, complejas y reducen el rendimiento. La tensión entre la innovación tecnológica acelerada y los marcos legales diseñados para paradigmas anteriores permanece, por ahora, irresuelta.

104 WIPO (2025).
105 Getty no obtuvo reconocimiento sustancial en el frente de copyright, pero logró un pronunciamiento limitado sobre infracción de marca.
106 Chatgptiseatingtheworld (2026).
107 Bartz v. Anthropic, Kadrey v. Meta, en California.
108 Thomson Reuters v. ROSS Intelligence, en Delaware.
109 AI Act (2024).

05 | Gobierno de la IA e impacto en personas

«Nuestra tecnología refleja nuestros valores».

Fei-Fei Li[110]

La adopción acelerada de IA plantea un desafío que trasciende lo tecnológico: cómo gobernar sistemas que evolucionan más rápido que las estructuras organizativas, cómo transformar roles profesionales cuando las tareas cambian trimestralmente, y cómo preservar el juicio humano en decisiones críticas mientras se delega la operación rutinaria a sistemas autónomos.

Este bloque aborda la dimensión organizativa y humana de la IA: los modelos de gobierno corporativo que están emergiendo para controlar la IA de extremo a extremo, las prácticas operativas avanzadas que permiten industrializar su despliegue (MLOps, LLMOps), la transformación de roles y perfiles profesionales que ya está en marcha, la adopción sectorial que convierte la IA en infraestructura transversal, el impacto en la vida cotidiana de las personas más allá del trabajo, los desafíos de sostenibilidad e impacto social que introduce, y los marcos éticos operativos que deben traducir principios abstractos en controles concretos y auditoría continua. La pregunta no es si la IA transformará las organizaciones y las personas, sino si seremos capaces de gobernar esa transformación con rigor, velocidad y responsabilidad.

110 Fei-Fei Li (n. 1976), científica informática chino-estadounidense, pionera de la visión por computador, co-directora del Stanford Human-Centered AI Institute (HAI). Conocida como «madrina de la IA», es una de las principales defensoras de una IA centrada en el ser humano.

Gobierno corporativo de la IA

La IA desborda los marcos tradicionales

Los modelos de gobierno corporativo tradicionales fueron diseñados para tecnologías predecibles: sistemas que ejecutan lógica determinista, operan dentro de límites definidos y se comportan de forma reproducible. La IA rompe todas estas premisas: toma decisiones sin intervención humana, produce *outputs* diferentes ante el mismo *input*, opera mediante procesos internos opacos que sus propios desarrolladores no comprenden por completo, y depende críticamente de proveedores externos cuyos modelos evolucionan sin control directo de la organización.

Las estructuras tradicionales de gobernanza tecnológica resultan insuficientes. Los comités de arquitectura y los ciclos de aprobación anuales son demasiado lentos para la velocidad de innovación en IA, carecen del *expertise* necesario para evaluar sus riesgos específicos, y no están diseñados para gestionar la incertidumbre inherente a sistemas que aprenden y cambian. La pregunta estratégica es cómo gobernar la IA sin frenar la velocidad de adopción y sin acumular riesgo ingobernable.

Organización: roles y estructuras que emergen

Las organizaciones están creando funciones ejecutivas especializadas, aunque con avance heterogéneo. Emerge la figura del *Chief AI Officer* (CAIO), *Chief Data and AI Officer* (CDAIO), o equivalente, y se estima[111] que un 26 % de las grandes organizaciones ya disponen de este rol[112]. En organizaciones más maduras, el CDAIO reporta directamente a presidencia, con una estructura matricial muy fuerte en países y negocios; en otras, el liderazgo recae en el CTO/CIO o *Chief Innovation Officer*.

Por debajo del nivel ejecutivo emergen roles especializados todavía sin estandarización: *AI Risk Manager, AI Ethics Officer, AI Compliance Lead* o *Responsible AI Lead.* Estos perfiles suelen reportar a las funciones de segunda línea, pero sus responsabilidades, autoridad y recursos varían sustancialmente entre organizaciones.

En primera línea se consolida el Centro de Excelencia de IA, equipos transversales que centralizan conocimiento técnico (LLMs, arquitecturas multiagente, *frameworks*), infraestructura compartida (plataformas *cloud*, GPUs, licencias), emisión de guías y estándares, y consultoría interna a líneas de negocio. Operativamente, domina el modelo *hub & spokes*: un Centro de Excelencia central establece capacidades transversales, mientras equipos descentralizados en líneas de negocio desarrollan soluciones específicas, reportando jerárquicamente a sus superiores, pero con reporte funcional cruzado al *hub*. Esta estructura permite velocidades diferentes según la madurez de cada línea manteniendo coherencia arquitectónica.

En segunda línea, la madurez organizativa es notablemente menor. Mientras la primera línea ha consolidado estructuras, la segunda presenta alta variabilidad, y hay un debate conceptual abierto sobre si *AI Risk* constituye un riesgo autónomo en la taxonomía corporativa o un amplificador transversal de riesgos existentes. Muy pocas organizaciones lo definen como riesgo de primer nivel (como decisión táctica para asegurar visibilidad y presupuesto); más frecuentemente, aparece como riesgo de segundo nivel dentro de Riesgo de Modelo o Riesgos No Financieros. Otras rechazan incluirlo formalmente y tratan la IA como un amplificador transversal de riesgos existentes (riesgo de modelo aumentado, riesgo de proveedor aumentado, riesgo tecnológico con vulnerabilidades adicionales…), reforzando marcos ya vigentes en lugar de crear taxonomías nuevas.

Independientemente del debate taxonómico, emerge operativamente una función pequeña de coordinación, llamada *AI Risk* o *AI Governance*, que orquesta las evaluaciones de riesgo por parte de las funciones especializadas: Riesgo de Modelo valida los modelos, Riesgo Tecnológico evalúa la infraestructura, Ciberseguridad analiza las vulnerabilidades específicas de la IA, *Data Protection* verifica la privacidad, Legal evalúa la propiedad intelectual, *Compliance* verifica el cumplimiento regulatorio, etc.

Órganos de gobierno: del comité formal al working group operativo

Prácticamente todas las organizaciones sistémicas han constituido un Comité de Inteligencia Artificial con composición cross-línea de defensa: primera línea (innovación, *analytics*, datos, tecnología…), segunda línea (riesgo de modelo, riesgo operacional, ciberseguridad, protección de datos, *vendor risk*, legal, *compliance*…), y tercera línea (Auditoría como observador). La co-presidencia suele recaer en un responsable de primera y uno de segunda línea.

Sin embargo, el gobierno no ocurre solo en el comité. Por debajo suele existir una estructura informal pero crítica: un *AI Working Group* compuesto por quienes reportan directamente a los miembros del comité. Preparan materiales, consensúan posiciones, resuelven conflictos. Los temas llegan al comité «*for information*» o «*for approval*», no para debate desde cero. El verdadero gobierno (la negociación, el consenso, la resolución de tensiones entre velocidad y control) ocurre en esta capa pre-decisional donde primera y segunda línea construyen acuerdos operativos antes de la sanción formal.

Arquitectura de marcos de riesgo: columna vertebral y uplifting sectorial

Las organizaciones no reinventan sus marcos de riesgo desde cero. Se observa una arquitectura de dos niveles: una columna vertebral central específica de IA (una política de IA breve, que establece principios, alcance, roles, clasificación de riesgos y procesos de aprobación; y un procedimiento de IA que describe exhaustivamente el ciclo de vida desde ideación hasta producción), y sobre todo el *uplifting* de marcos específicos existentes.

El *uplifting* consiste en complementar los marcos vigentes añadiendo capítulos específicos de IA. El marco de Riesgo de Modelo incorpora validación de IA generativa, técnicas de explicabilidad (SHAP, LIME, *attention mechanisms*), detección de sesgos y alucinaciones, monitorización de *drift*, etc. El marco de Riesgo de Proveedor añade *due diligence* sobre proveedores de LLMs, certificaciones requeridas, SLAs específicos, planes de contingencia o *exit strategies*. El marco de Protección de Datos incluye evaluaciones de impacto específicas para IA, minimización

111 IBM (2025b).

112 Entre ellas, JPMorganChase, Santander, Société Générale, Rabobank, Scotiabank, Visa, Mastercard, AXA, Pfizer, Merck, S&P, Microsoft, LinkedIn, eBay, T-Mobile, Orange, Schneider Electric, SAP, Marks and Spencer, Boeing, Nike, Michelin, por citar algunas. Ver PixieBrix (2025).

de datos en sistemas de IA o ejercicio de derechos GDPR cuando el procesamiento involucra IA. El marco de *Compliance* implementa el *AI Act*: clasificación de riesgo, registro, documentación, etc.

Esta arquitectura aprovecha la infraestructura que funciona, asigna responsabilidades claras y mantiene coherencia sin fragmentación. Sin embargo, algunos desafíos técnicos son genuinamente nuevos. Explicar redes neuronales profundas con *transformers* y trillones de parámetros requiere técnicas especializadas que no existían en la validación de modelos tradicionales. Las funciones de Riesgos están desarrollando estas capacidades en tiempo real.

Clasificación de riesgos y ciclo de vida

El *AI Act* europeo establece una clasificación (prohibido, alto riesgo, riesgo limitado, riesgo mínimo), necesaria pero insuficiente para el gobierno corporativo de las compañías. La regulación está concebida para proteger a la sociedad y los derechos fundamentales, no a las organizaciones frente sus propios riesgos operacionales, reputacionales o financieros.

Por ello, las organizaciones convergen hacia clasificaciones de riesgo de la IA internas más exigentes, que integran criterios regulatorios con propios: impacto reputacional, criticidad del proceso, clasificación de ciberseguridad, *tier* del modelo según validación, impacto financiero directo, madurez de proveedores, etc. Un sistema puede no ser alto riesgo regulatorio, pero sí serlo internamente si afecta a procesos críticos o genera una exposición reputacional masiva (Fig. 4).

La clasificación determina las consecuencias operativas. Por ejemplo, alto riesgo implica un análisis exhaustivo de todas las funciones de riesgo, la aprobación formal en Comité, documentación completa, validación independiente, monitorización intensiva, etc.; mientras que bajo riesgo suele implicar un *fast track,* análisis ligero y elevación al Comité solo para información. El *fast track* es un elemento crítico, porque resuelve la tensión entre velocidad y control: los sistemas de IA evidentemente de bajo riesgo se aprueban sin demoras, mientras se concentran los recursos de control en los sistemas de IA críticos.

Inventario y trazabilidad

Las organizaciones están obligadas por regulación a mantener un registro único de todos los sistemas de IA desplegados, incluyendo su clasificación de riesgos y estado en el ciclo de vida. Este inventario debe ser completo, actualizado y accesible para supervisores, auditores y funciones de riesgo.

La mayoría de las organizaciones converge hacia expandir el inventario de Riesgo de Modelo, bajo la lógica de que los sistemas de IA requieren validación cuando comportan un riesgo de modelo relevante. Alternativamente, algunas organizaciones expanden el inventario de activos tecnológicos.

El gobierno corporativo de la IA está en maduración acelerada. Los próximos años verán una consolidación hacia mejores prácticas, pero la aceleración de la evolución tecnológica es tal que probablemente continuará tensionando las estructuras organizativas actuales, que fueron diseñadas para paradigmas más lentos.

Fig. 4. Ejemplo de clasificación interna de riesgos de una compañía, más allá de la regulación.

Criterios de riesgo

Criterio	Disparador	Nivel
Prohibido por regulación	Sistema prohibido / restricción aplicable	PROHIBIDO
Alto riesgo por regulación	Sistema de alto riesgo por norma aplicable	ALTO
Proceso	Impacta proceso crítico de la empresa	ALTO
Personas	Decisiones relevantes sobre personas físicas	ALTO
Autonomía	Actúa autónomamente en nombre de la firma	ALTO
Reputación	Interacción directa con externos	ALTO
Legal	Efectos legales o contractuales relevantes	ALTO
Privacidad	Datos personales	ALTO
Confidencialidad	Datos confidenciales de la empresa o clientes	MEDIO
Terceros	Dependencia crítica no sustituible	MEDIO
Resto	Sin disparadores anteriores	BAJO

+

Mitigantes

Mitigante	Disparador
Human in the loop	Revisión humana del 100% de outputs
Auditoría	Auditoría / evaluación técnica periódica
Apelación	Mecanismo de apelación accesible para personas sujetas al sistema
Anonimización	Datos personales/confidenciales anonimizados o pseudonimizados

=

Riesgo del sistema de IA

- **Riesgo inherente**: el nivel más alto entre los criterios de riesgo.
- **Riesgo residual**: el nivel más alto tras aplicar mitigantes, criterio a criterio

PROHIBIDO | ALTO | MEDIO | BAJO

Industrialización de la IA (MLOps, LLMOps)

De la experimentación a la producción

En el momento actual, el principal cuello de botella en la adopción real de la IA no es algorítmico, sino operativo. Durante años, organizaciones de todo tipo han desarrollado sistemas de IA prometedores en entornos experimentales que nunca llegaron a producción, o que, una vez desplegados, fallaron al enfrentarse a datos reales, perdieron rendimiento con el tiempo o generaron costes y riesgos inasumibles. La industrialización de la IA surge precisamente para cerrar esta brecha entre la experimentación y el uso sostenido en entornos productivos.

Machine Learning Operations (MLOps) emerge como respuesta estructurada a este problema. Puede entenderse como «un conjunto de procesos estandarizados y capacidades tecnológicas para construir, desplegar y operacionalizar sistemas de ML de forma rápida y confiable»[113]. MLOps articula procesos y capacidades técnicas para gestionar de forma integrada la preparación de datos, la experimentación, el entrenamiento, la validación, el despliegue, la monitorización y el reentrenamiento continuo de modelos[114]. El objetivo no es únicamente acelerar la puesta en producción, sino garantizar fiabilidad, reproducibilidad, control de riesgos y sostenibilidad operativa a lo largo del tiempo.

La irrupción de la IA generativa amplía este reto de forma significativa. *Large Language Model Operations* (LLMOps) no sustituye a MLOps, sino que lo extiende para gestionar sistemas con propiedades radicalmente distintas[115]. Los grandes modelos de lenguaje introducen comportamiento no determinista, dependencia crítica de la formulación de *prompts*, arquitecturas internas opacas con miles de millones o billones de parámetros, y nuevos vectores de riesgo como alucinaciones, generación de contenido no veraz o ataques específicos mediante manipulación de entradas. Estas complejidades se intensifican en sistemas agénticos, donde una única interacción del usuario puede desencadenar cadenas de razonamiento, múltiples llamadas internas al modelo y ejecución autónoma de herramientas externas.

Arquitectura del ciclo de vida MLOps/LLMOps

La industrialización efectiva de la IA requiere una visión integral del ciclo de vida operativo, que diferentes autores expresan de distintos modos (Fig. 5), pero con elementos comunes:

En la fase de **preparación de datos**, MLOps se centra en la construcción de *pipelines* robustos para la ingesta, limpieza, transformación y versionado de datos estructurados, garantizando trazabilidad y calidad. LLMOps amplía este alcance al trabajar con grandes volúmenes de datos no estructurados (texto, código, documentos o imágenes) y con representaciones vectoriales utilizadas en arquitecturas de *retrieval-augmented generation* (RAG). A ello se suman requisitos estrictos de minimización de datos personales, control de fuentes y cumplimiento de marcos de privacidad como GDPR.

Durante la **experimentación y el desarrollo**, MLOps proporciona mecanismos para rastrear experimentos, comparar modelos, gestionar hiperparámetros y asegurar la reproducibilidad de resultados. En LLMOps, esta fase incorpora nuevas dimensiones: gestión versionada de *prompts*, evaluación de distintas configuraciones de modelos fundacionales, fine-tuning o adaptación mediante técnicas como LoRA, y diseño de métricas que capturen aspectos cualitativos de la generación, como coherencia, factualidad, seguridad o adecuación al contexto. Estas métricas no sustituyen a las cuantitativas tradicionales, sino que las complementan allí donde el rendimiento ya no puede medirse únicamente con precisión o error.

La **validación** constituye uno de los puntos de mayor divergencia entre ambos enfoques. Mientras que en MLOps la validación se apoya en test automatizados sobre conjuntos de datos de referencia para evaluar precisión, sesgo y robustez, en LLMOps resulta imprescindible integrar validación humana en el *loop*. La evaluación de salidas generativas exige revisión experta, técnicas de *fact-checking* contra fuentes confiables, pruebas de estrés semántico y ejercicios de *red-teaming* diseñados para identificar comportamientos no deseados o vulnerabilidades explotables. La validación deja de ser un evento puntual para convertirse en un proceso continuo.

En la fase de **despliegue**, MLOps se apoya en *pipelines* CI/CD relativamente maduros, con versiones bien definidas de modelos y dependencias. LLMOps, en cambio, debe gestionar despliegues de modelos de gran escala con implicaciones significativas de infraestructura, latencia y coste. Esto incluye selección dinámica de modelos en función del caso de uso, del nivel de riesgo o del presupuesto, así como la orquestación de sistemas complejos en los que múltiples modelos y agentes interactúan entre sí de forma coordinada.

La **monitorización** es crítica en ambos paradigmas, pero con énfasis distintos. MLOps se centra en detectar *drift* de datos o de concepto, degradación de rendimiento, errores y problemas de latencia. LLMOps añade la necesidad de monitorizar costes por *token* en tiempo real (un factor determinante para la viabilidad económica), identificar patrones de uso no previstos (*prompt* drift), y mantener trazabilidad completa de interacciones para análisis forense, auditoría y supervisión regulatoria. Sin esta visibilidad, los sistemas generativos pueden escalar rápidamente en complejidad y coste sin un control efectivo.

Finalmente, la **gobernanza y el cumplimiento normativo** atraviesan todo el ciclo de vida. En MLOps, esto implica documentar el linaje de datos y modelos, definir responsabilidades claras y asegurar controles de calidad. En LLMOps, estas exigencias se amplían para dar respuesta a marcos regulatorios emergentes como el *AI Act*, incluyendo inventarios de sistemas clasificados por nivel de riesgo, mecanismos de supervisión humana y estrategias de explicabilidad adaptadas a modelos que funcionan como cajas negras.

113 Google (2023).
114 iDanae (3Q20).
115 Shan (2024).

Integración con marcos de gobernanza

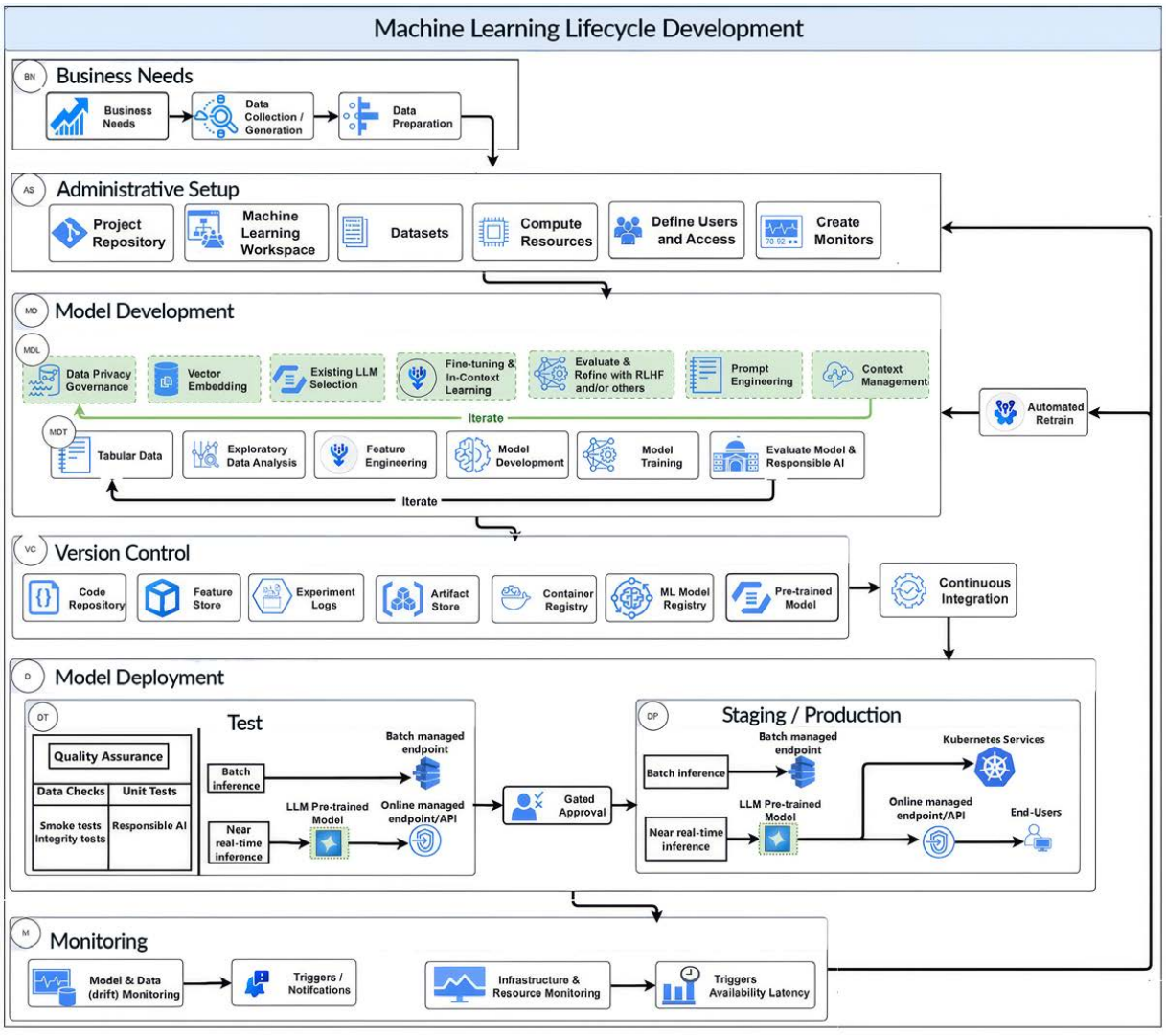
MLOps y LLMOps no son disciplinas puramente técnicas ni pueden operar de forma aislada. Constituyen la capa operativa que materializa los principios definidos en marcos de gobernanza más amplios. El NIST *AI Risk Management Framework*[116] estructura la gestión del riesgo de IA como un proceso continuo que abarca gobierno, contextualización, medición y gestión a lo largo de todo el ciclo de vida del sistema. De forma complementaria, ISO/IEC 42001 define[117] los requisitos de un sistema de gestión de IA, incluyendo políticas, evaluaciones de impacto, control de proveedores y supervisión continua.

Los procesos automatizados de MLOps y LLMOps (*pipelines* de datos, validaciones sistemáticas, despliegues controlados y monitorización permanente) son los mecanismos concretos que permiten que estos marcos de gobernanza pasen de la teoría a la práctica. Sin una base operativa sólida, la gobernanza de la IA queda reducida a documentación aspiracional; sin marcos de gobernanza claros, la industrialización técnica carece de criterios de calidad, responsabilidad y control del riesgo.

La adopción madura de la IA exige, por tanto, una convergencia real entre ingeniería, operaciones y gobernanza. La industrialización de la IA no consiste únicamente en escalar modelos, sino en construir sistemas confiables, auditables y sostenibles que puedan integrarse de forma responsable en procesos críticos de negocio. MLOps y LLMOps son marcos de buenas prácticas en evolución activa, no soluciones cerradas: si el problema de llevar IA a producción estuviera resuelto, no se observaría la persistencia de modelos que se degradan silenciosamente, generan costes no anticipados o fallan al escalar. Su valor es reducir la fricción y estructurar la gestión del riesgo, no eliminarlo.

116 NIST (2023).
117 ISO/IEC (2023).

Fig. 5. Ejemplo de fases de MLOps y LLMOps. Fuente: Stone (2025).



Upskilling, *reskilling* y nuevos roles profesionales

El factor humano, cuello de botella de la IA

La proliferación de la IA está transformando el trabajo de forma más profunda de lo que sugieren los debates centrados únicamente en la automatización o sustitución de tareas. El principal reto para las organizaciones no es tecnológico, sino humano: disponer de las capacidades adecuadas para diseñar, desplegar, operar y gobernar sistemas de IA de manera sostenible. En este contexto, el talento deja de entenderse como un conjunto reducido de expertos y pasa a concebirse como una competencia organizativa distribuida.

Las conclusiones de esta sección se apoyan en un análisis empírico exhaustivo realizado por Management Solutions sobre un conjunto de 16 grandes organizaciones de Europa y Estados Unidos, basado en datos reales de mercado y evidencia organizativa.

Convergencia hacia un núcleo estable de roles de IA

A medida que la IA madura, las organizaciones convergen hacia un conjunto relativamente estable de competencias y capacidades profesionales. Aunque la nomenclatura varía entre sectores y empresas, las funciones que desempeñan estos perfiles comienzan a ser homogéneas. Sin embargo, los roles formales son menos que los *skillsets* subyacentes: en la práctica, un mismo profesional combina varias de estas capacidades, y las combinaciones más inusuales (quien domina a la vez LLMOps, regulación y *prompt engineering*, por ejemplo) son precisamente las más valiosas y escasas. Esta convergencia refleja una realidad operativa: el ciclo de vida de la IA (desde los datos hasta la producción y el control) requiere capacidades diferenciadas que no pueden concentrarse en un único tipo de profesional.

De forma sintética, estos roles pueden agruparse en tres grandes bloques (Fig. 6). En primer lugar, perfiles técnicos, responsables de diseñar modelos, construir infraestructuras de datos y llevar soluciones a producción. En segundo lugar, perfiles híbridos, que conectan capacidades técnicas con necesidades de negocio y aseguran que los sistemas de IA generen valor real y sean adoptados. Por último, perfiles de gobierno y control, encargados de gestionar riesgos, cumplimiento normativo, auditoría y uso responsable de la IA.

Esta estructura de competencias constituye la columna vertebral sobre la que se construyen las capacidades de IA de las organizaciones, independientemente del sector en el que operen.

Madurez desigual en la adopción de roles especializados

Aunque el núcleo básico de roles de IA está ampliamente implantado en organizaciones avanzadas (con heterogeneidad significativa), la adopción de perfiles emergentes o más especializados presenta una madurez muy desigual. Las diferencias no se observan tanto en la existencia de capacidades generales, como en la institucionalización de funciones avanzadas relacionadas con la operación de IA generativa y agéntica, la arquitectura a gran escala o el gobierno específico de estos sistemas.

Esta heterogeneidad es especialmente visible en roles emergentes vinculados a LLMOps, evaluación continua de modelos generativos, seguridad de *prompts* o gobierno de IA. En algunos casos, estas funciones existen como roles formales; en otros, se integran de manera informal en equipos más tradicionales, con resultados dispares en términos de control, escalabilidad y coste.

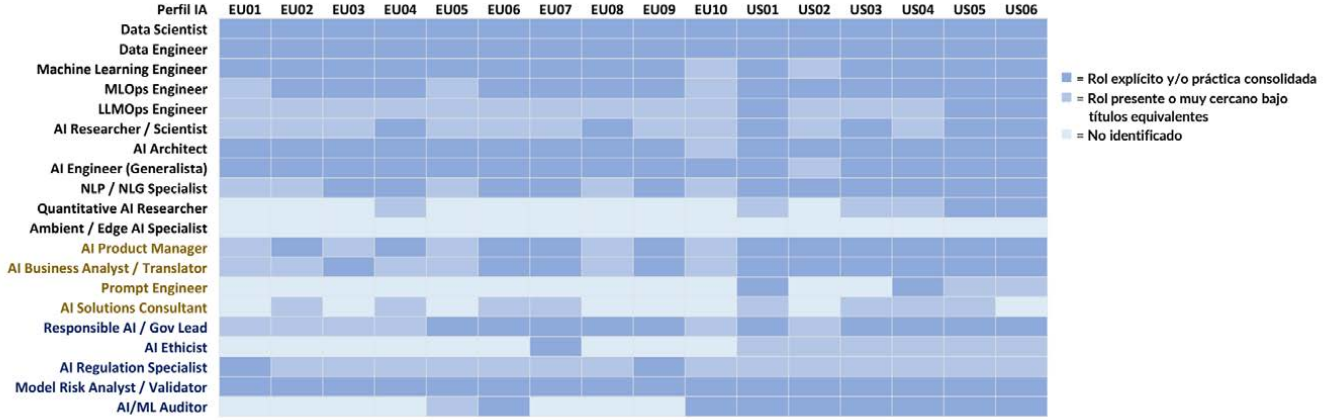
En el mapa de calor[118] (Fig. 7) se ilustra que la diferencia entre organizaciones se manifiesta en qué roles han consolidado y con qué nivel de especialización y autonomía.

118 El mapa de calor se basa en un análisis comparativo de la adopción de roles de IA en un conjunto de 16 grandes organizaciones europeas y estadounidenses. La identificación de roles se ha realizado a partir de fuentes públicas (ofertas de empleo, estructuras organizativas o iniciativas estratégicas anunciadas) y evidencia cualitativa de mercado, considerando tanto roles explícitos como funciones equivalentes bajo denominaciones alternativas. La intensidad del color refleja el grado de institucionalización del rol, desde ausencia del perfil hasta práctica consolidada.

Fig. 6. Perfiles consolidados y emergentes en IA.

#	Perfil IA	Tipo	Descripción
01	Data Scientist	Técnico	Diseña modelos predictivos y extrae insights accionables para mejorar productos y decisiones del banco.
02	Data Engineer	Técnico	Construye y mantiene la infraestructura y los pipelines de datos que alimentan modelos y analítica.
03	Machine Learning Engineer	Técnico	Despliega modelos en producción y desarrolla APIs/servicios para que el negocio los consuma de forma segura y eficiente.
04	MLOps Engineer	Técnico	Automatiza y opera el ciclo de vida de los modelos (CI/CD, monitorización, retraining) garantizando su robustez.
05	LLMOps Engineer	Técnico	Industrializa GenAI en producción (routing, caching, observabilidad, evaluación continua, guardrails, despliegue y coste).
06	AI Researcher / Scientist	Técnico	Investiga nuevas técnicas de IA/ML y desarrolla prototipos avanzados que puedan convertirse en capacidades del banco.
07	AI Architect	Técnico	Diseña la arquitectura de soluciones de IA, integrando modelos con sistemas core y asegurando escalabilidad y seguridad.
08	AI Engineer (Generalist)	Técnico	Construye soluciones de software que integran componentes de IA pre-entrenados dentro de aplicaciones del banco.
09	NLP / NLG Specialist	Técnico	Desarrolla modelos de lenguaje y soluciones basadas en LLMs para analizar, resumir y generar contenido textual.
10	Quantitative AI Researcher	Técnico	Aplica IA y deep learning a mercados y estrategias cuantitativas para mejorar señales, predicciones y trading.
11	Ambient / Edge AI Specialist	Técnico	Despliega IA en entornos IoT/edge y sistemas "ambient" (sensores, dispositivos), integrando datos en tiempo real.
12	AI Product Manager	Híbrido	Lidera el ciclo de vida de productos de IA, traduce capacidades técnicas en valor de negocio y prioriza su evolución.
13	AI Business Analyst / Translator	Híbrido	Identifica casos de uso de IA y conecta necesidades del negocio con requisitos técnicos accionables.
14	Prompt Engineer	Híbrido	Diseña, prueba y estandariza prompts para casos de uso GenAI, y protege apps GenAI contra prompt injection y jailbreaks.
15	AI Solutions Consultant	Híbrido	Lidera la implantación de soluciones de IA en áreas de negocio, asegurando valor real y adopción efectiva.
16	Responsible AI / AI Governance Lead	Gobierno	Define políticas de IA responsable y supervisa el cumplimiento ético, regulatorio y de riesgo de los modelos.
17	AI Ethicist	Gobierno	Define y operacionaliza criterios éticos (sesgos, impacto social...) en casos de uso de IA, con guardrails y revisión de decisiones.
18	AI Regulation Specialist	Gobierno	Traduce regulación (AI Act y análogos), políticas internas y expectativas supervisoras en requisitos, controles, evidencias y reporting.
19	Model Risk Analyst / Validator	Gobierno	Valida modelos de IA/ML mediante análisis independiente para asegurar robustez, equidad y cumplimiento normativo.
20	AI/ML Auditor	Gobierno	Audita procesos, controles y uso de IA para verificar cumplimiento de políticas internas y regulaciones.

Fig. 7. Mapa de calor de adopción de roles de IA.



El cuello de botella real: desequilibrios oferta–demanda

El análisis de oferta y demanda[119] (Fig. 8) muestra un desequilibrio estructural generalizado en el mercado de talento en IA. La demanda es elevada en prácticamente todos los perfiles analizados, mientras que la oferta resulta insuficiente de forma sistemática para absorberla. No se trata de una escasez puntual o concentrada, sino de un fenómeno transversal que afecta a la mayor parte del ecosistema profesional de la IA.

La única excepción parcial es el perfil de *Data Scientist*, que presenta una situación más cercana al equilibrio relativo. Este comportamiento responde a su mayor madurez histórica, a la existencia de trayectorias formativas consolidadas y a un mayor *pool* de profesionales con experiencia acumulada. Sin embargo, incluso en este caso, la presión de la demanda se mantiene elevada en posiciones senior o especializadas.

En el resto de perfiles, especialmente aquellos orientados a producción (*Machine Learning Engineering*, MLOps, LLMOps) y a funciones de gobierno, riesgo y control, la brecha entre demanda y oferta es persistente. La combinación de complejidad técnica, *seniority* requerida y creciente exigencia regulatoria limita la capacidad del mercado para generar talento al ritmo necesario. Como resultado, la contratación externa, por sí sola, no permite cerrar el gap, convirtiendo el *upskilling* y *reskilling* internos en una palanca estructural para escalar la IA con impacto y control.

Implicaciones estratégicas

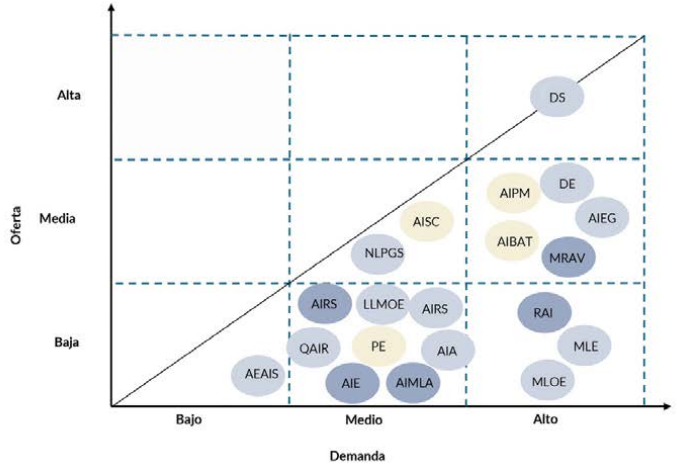
Las implicaciones de este análisis son claras. La estrategia de talento en IA no puede basarse exclusivamente en la captación de perfiles escasos en el mercado. El *upskilling* y *reskilling* internos se convierten en palancas estratégicas, especialmente para roles críticos donde la oferta externa es estructuralmente insuficiente.

Asimismo, los perfiles de operación y gobierno son tan determinantes como los de modelado o investigación. La ventaja competitiva no reside únicamente en desarrollar modelos avanzados, sino en integrarlos de forma segura, eficiente y conforme a regulación en los procesos reales de la organización.

En última instancia, la transformación impulsada por la IA no es solo tecnológica. Es una transformación del trabajo, de los roles profesionales y de las capacidades colectivas. Las organizaciones que entiendan esta dimensión humana y la aborden de forma estructurada estarán mejor posicionadas para capturar el valor de la IA de manera sostenida.

119 La estimación de la demanda se basa en un análisis longitudinal de ofertas de empleo publicadas entre el segundo semestre de 2025 y enero de 2026 por grandes organizaciones europeas y estadounidenses, considerando volumen relativo por perfil, seniority exigida y recurrencia temporal. Este análisis se complementa con informes sectoriales de mercado laboral y evidencia cualitativa de iniciativas estratégicas en IA (creación de centros de excelencia, funciones de IA responsable, riesgo de modelos y GenAI).
La oferta se estima de forma relativa a partir de la dificultad de cobertura observada en el mercado, incluyendo primas salariales, seniority requerida y escasez reportada en estudios especializados. Adicionalmente, se incorpora la madurez histórica de cada perfil (antigüedad de la disciplina, existencia de pipelines formativos) y la viabilidad de transición desde roles adyacentes mediante upskilling o reskilling, contrastada con observaciones de mercado sobre la disponibilidad efectiva de profesionales con experiencia demostrable.

Fig. 8. Matriz oferta–demanda de perfiles de IA



Siglas	Perfil IA
DS	Data Scientist
DE	Data Engineer
MLE	Machine Learning Engineer
MLOE	MLOps Engineer
LLMOE	LLMOps Engineer
AIRS	AI Researcher / Scientist
AIA	AI Architect
AIEG	AI Engineer (Generalist)
NLPGS	NLP / NLG Specialist
QAIR	Quantitative AI Researcher
AEAIS	Ambient / Edge AI Specialist
AIPM	AI Product Manager
AIBAT	AI Business Analyst / Translator
PE	Prompt Engineer
AISC	AI Solutions Consultant
RAI	Responsible AI / AI Governance Lead
AIE	AI Ethicist
AIRS	AI Regulation Specialist
MRAV	Model Risk Analyst / Validator
AIMLA	AI/ML Auditor

IA y transformación sectorial (AI + X)

La IA como capa transversal

La IA ha dejado de ser una tecnología aplicada de forma incremental en sectores específicos para convertirse en una capa transversal de inteligencia que se integra simultáneamente en múltiples dominios económicos y sociales. Esta transición, ampliamente documentada por organismos internacionales, se caracteriza por el paso desde pilotos aislados hacia una adopción estructural que afecta a procesos completos, cadenas de valor y modelos operativos. No todo el avance responde a una lógica transversal: algunos de los desarrollos más significativos son impulsados por sectores específicos (la IA médica, la defensa o los servicios financieros regulados) donde la especificidad del dominio, la naturaleza de los datos o el marco regulatorio generan capacidades que difícilmente se transfieren a otros contextos sin adaptación sustancial.

Desde una perspectiva comparada, la evidencia muestra que las diferencias entre sectores ya no se explican por la mera presencia de IA, sino por la intensidad y profundidad de su integración. La OCDE[120] clasifica los sectores según su *AI intensity* atendiendo a factores como la composición de tareas, la disponibilidad de datos y el capital humano, mostrando que incluso sectores tradicionalmente poco digitalizados están incrementando rápidamente su exposición a la IA. Esta adopción ocurre de forma paralela entre sectores, generando efectos de aceleración cruzada y aprendizaje intersectorial[121].

En términos macroeconómicos y laborales, el impacto es sistémico. El Fondo Monetario Internacional estima[122] que alrededor del 40 % del empleo global está expuesto a la IA, con porcentajes superiores en economías avanzadas, donde la complementariedad entre IA y trabajo cualificado es mayor. La Organización Internacional del Trabajo matiza[123] que la IA generativa tiende a automatizar tareas concretas más que ocupaciones completas, reforzando la necesidad de adaptación organizativa y formación continua.

En este contexto, el concepto AI + X resulta útil para describir un patrón estructural: la IA se combina con dominios sectoriales específicos sin perder su carácter de tecnología de propósito general. No se trata de una suma de casos de uso independientes, sino de una integración simultánea y transversal, respaldada por evidencia empírica reciente[124]. En esta sección se presentan algunos ejemplos ilustrativos, sin ánimo de exhaustividad (Fig. 9).

Salud (AI + Health): precisión clínica y escalabilidad

En el ámbito sanitario, la IA se está integrando en diagnóstico clínico, apoyo a la decisión médica, monitorización de pacientes y automatización administrativa. La Organización Mundial de la Salud documenta[125] un crecimiento sostenido del uso de sistemas de IA en radiología, anatomía patológica y atención primaria, destacando tanto su potencial clínico como los riesgos asociados.

En múltiples estudios[126] se demuestra que los sistemas de IA alcanzan niveles de precisión comparables o superiores a especialistas humanos en tareas como detección de cáncer en imágenes médicas o clasificación dermatológica. Estos resultados no implican sustitución directa, pero sí una ampliación significativa de la capacidad diagnóstica y de cribado a gran escala. Este avance, subraya la OMS, exige marcos sólidos de gobernanza, validación clínica y supervisión humana, dado su impacto directo en personas.

Educación (AI + Education): personalización a escala

En educación, la IA generativa está transformando la creación de contenidos, la evaluación y el acompañamiento al aprendizaje. UNESCO identifica[127] el uso creciente de tutores adaptativos, generación automática de materiales didácticos y sistemas de evaluación formativa continua, con potencial para mejorar la personalización y reducir brechas educativas.

La educación[128] es uno de los sectores donde la IA puede tener impactos estructurales a medio plazo, al permitir modelos de aprendizaje más flexibles y adaptativos. No obstante, estos beneficios dependen críticamente de políticas públicas y marcos institucionales que preserven la equidad, la integridad académica y el rol del profesorado.

Trabajo y empleo (AI + Work): reconfiguración de tareas

La IA está redefiniendo el trabajo de forma transversal, afectando a ocupaciones en prácticamente todos los sectores. El IMF estima[129] que la exposición a la IA alcanza aproximadamente el 60 % del empleo en economías avanzadas, frente a porcentajes menores en economías emergentes, reflejando diferencias en estructura productiva y capital humano.

La ILO, por su parte, concluye[130] que la IA generativa tiene por ahora un mayor impacto sobre tareas cognitivas específicas que sobre empleos completos, lo que implica una reconfiguración del contenido del trabajo más que una sustitución masiva. Esta evidencia refuerza la necesidad de estrategias de *upskilling* y *reskilling*, como se ha analizado anteriormente.

Industria y operaciones (AI + Industry): productividad y eficiencia

En entornos industriales y operativos, la IA se aplica a mantenimiento predictivo, optimización de procesos, planificación y robótica avanzada. Muchas organizaciones[131] han pasado ya de pruebas piloto a despliegues a escala, integrando la IA en procesos críticos de producción y logística.

Ya hay evidencia sólida[132] de mejoras significativas de eficiencia y reducción de fallos en sistemas industriales donde la IA se integra de forma sistemática, aunque los mayores beneficios se observan cuando la tecnología se acompaña de cambios organizativos y de procesos.

120 OECD (2024).
121 World Economic Forum (2025a).
122 IMF (2024).
123 ILO (2025a).
124 Stanford (2025).
125 WHO (2024).
126 Stanford (2025).
127 UNESCO (2023).
128 World Economic Forum (2025b).
129 IMF (2024).
130 ILO (2025a).
131 World Economic Forum (2025a).
132 Stanford HAI (2025).

Finanzas y servicios intensivos en información (AI + Finance)

En sectores intensivos en información, como finanzas y servicios profesionales, la IA se utiliza para detección de fraude, gestión de riesgos, personalización de servicios y automatización documental. Se observa[133] una expansión sostenida del uso de IA en estos ámbitos, con diferencias relevantes según el marco regulatorio y la capacidad organizativa.

La adopción de IA generativa en estos sectores se asocia a mejoras de productividad y a la aparición de nuevos modelos de servicio, aunque se advierte[134] que los beneficios dependen de la calidad de los datos, la gobernanza y la integración en procesos existentes.

Creatividad y contenidos (AI + Creative): nuevos modelos culturales

La generación de texto, imagen, música y vídeo mediante IA está transformando industrias creativas y de medios. Se observa[135] un crecimiento acelerado de estas aplicaciones y un aumento significativo de su uso comercial y cultural en los últimos años.

Este avance plantea debates regulatorios y económicos relevantes, especialmente en torno a propiedad intelectual y modelos de negocio, que están siendo abordados de forma desigual por distintos marcos normativos.

Síntesis: de sectores a infraestructura transversal

Más allá de los ejemplos sectoriales, la evidencia converge en un punto central: la transformación actual no reside en la adopción aislada de IA en sectores concretos, sino en su integración simultánea y transversal como infraestructura operativa y cognitiva. Los estudios coinciden en que la ventaja competitiva ya no se obtiene aplicando IA a funciones individuales, sino integrándola de forma coherente a lo largo de toda la cadena de valor.

En este sentido, AI + X no describe ya una moda terminológica, sino un patrón estructural de transformación. Comprender esta lógica resulta esencial para anticipar impactos sectoriales, diseñar políticas públicas adecuadas y definir estrategias organizativas capaces de capturar el valor de la IA de manera sostenible.

133 OECD (2025a); OECD (2026).
134 OECD (2025b).
135 Stanford (2025).

Fig. 9. Ejemplos representativos de AI + X, priorizando casos donde la IA ha pasado de piloto a uso operativo medible. Fuente: adaptado de Stanford (2025).

Dominio (X)	Ejemplo AI + X	Impacto
Salud – diagnóstico	Sistemas de IA ya detectan cáncer en imágenes médicas con precisión superior a humanos.	Permite cribado masivo y diagnóstico temprano a escala imposible para equipos humanos.
Salud – práctica clínica	Médicos usan asistentes de IA que escuchan la consulta y redactan la historia clínica.	Se reducen decenas de minutos diarios de burocracia médica; más tiempo para el paciente, menos burnout.
Movilidad urbana	Robotaxis operan de forma continua en varias ciudades, con cientos de miles de viajes semanales.	El transporte autónomo deja de ser experimento y se convierte en servicio urbano real.
Logística	Camiones autónomos transportan mercancías a gran escala, con millones de kilómetros acumulados.	Reducción de costes, operación 24/7 y menor dependencia de escasez de conductores.
Industria	Robots industriales con IA se instalan a un ritmo récord, especialmente en Asia.	La automatización avanzada se convierte en ventaja competitiva estructural, no puntual.
Robótica avanzada	Modelos de IA permiten que robots aprendan tareas complejas con menos programación manual.	Se acelera radicalmente el desarrollo y despliegue de robots versátiles.
Ingeniería de software	Programadores utilizan IA para escribir, depurar y documentar código de forma sistemática.	El desarrollo de software se acelera y cambia el rol del programador, más diseñador que mecanógrafo.
Atención al cliente	Sistemas generativos gestionan interacciones completas con clientes en centros de contacto.	Aumenta la productividad y se escalan servicios sin crecimiento proporcional de personal.
Supply chain	IA optimiza inventarios, previsión de demanda y planificación logística en tiempo real.	Menos roturas de stock, menos sobreinventario, mayor eficiencia operativa.
Marketing y ventas	Generación automática de campañas, contenidos y segmentación personalizada a gran escala.	Incremento medible de ingresos y reducción del tiempo de lanzamiento al mercado.
Estrategia corporativa	Ejecutivos usan IA para analizar escenarios, documentos estratégicos y alternativas de decisión.	La planificación estratégica se apoya en análisis continuo, no en ejercicios anuales.
Finanzas	IA automatiza análisis financiero, detección de anomalías y soporte a decisiones de inversión.	Mayor velocidad y cobertura analítica sin aumentar proporcionalmente los equipos.
Legal	Revisión de contratos y búsqueda jurídica automatizadas con IA generativa.	Diligencias legales que antes tomaban semanas se reducen a horas o días.
Recursos humanos	IA apoya selección, evaluación y gestión de talento con análisis masivo de información.	Procesos más rápidos y consistentes, con riesgos que requieren gobernanza explícita.
Educación	Tutores de IA adaptan contenidos y ritmo al nivel de cada estudiante.	Aprendizaje más personalizado, incluso en entornos con alta ratio alumno-profesor.
Creatividad	IA genera música, imágenes y vídeo que se integran en procesos creativos profesionales.	Cambia el proceso creativo: de producción manual a co-creación humano-máquina.
Medios y comunicación	Generación automática de contenidos informativos, resúmenes y traducciones.	Escalado de producción editorial con nuevos retos de calidad y veracidad.
Ciberseguridad	IA detecta patrones de ataque y anomalías en tiempo real.	Mejora la capacidad defensiva frente a amenazas crecientes y automatizadas.
Operaciones internas	IA automatiza tareas transversales (documentación, reporting, análisis interno).	Reducción estructural de costes administrativos y fricción organizativa.
Economía digital	La mayoría de grandes organizaciones ya usan IA generativa en al menos una función crítica.	La IA deja de ser diferencial y pasa a ser infraestructura básica.

IA en la vida personal y cotidiana

La inversión del paradigma: primero las personas, después las organizaciones

La IA generativa ha protagonizado una inversión radical del paradigma tecnológico tradicional. A diferencia de tecnologías empresariales anteriores (*cloud computing*, ERP, CRM) que nacieron en entornos corporativos y gradualmente se filtraron hacia el consumo personal[136], la IA generativa irrumpió primero en la vida cotidiana de las personas y solo después fue adoptada formalmente por las organizaciones.

Los datos son contundentes: se estima[137] que ChatGPT alcanzó aproximadamente 900 millones de usuarios activos semanales a finales de 2025. Esta cifra coloca a la IA generativa entre las tecnologías de adopción más rápida de la historia, superando la velocidad de penetración de internet móvil, redes sociales o servicios de *streaming*.

Más revelador aún, la proporción de uso personal no solo precede, sino que supera sistemáticamente al uso profesional. En la Unión Europea[138], el 25,1 % de la población utiliza herramientas de IA generativa para propósitos personales, frente al 15,1 % que las emplea en contextos laborales y apenas el 9,4 % en educación formal. En los países de la OCDE[139], más de un tercio de los individuos reportan uso regular de IA generativa, con los estudiantes liderando la adopción: tres cuartas partes de estudiantes mayores de 16 años utilizan estas herramientas, mientras que el 41,1 % de los empleados las integra en su trabajo, frecuentemente de manera informal incluso antes de que sus organizaciones lo autoricen.

Esta secuencia temporal no es anecdótica: redefine la lógica de la transformación digital. Las organizaciones no están liderando la adopción de IA; están respondiendo a capacidades que sus empleados, clientes y proveedores ya poseen y utilizan extraoficialmente. El fenómeno del *shadow AI* (uso no autorizado de herramientas de IA en entornos corporativos) no es una anomalía transitoria, sino una consecuencia estructural de esta inversión: las personas adquirieron competencias cognitivas aumentadas en su vida personal que luego trasladaron espontáneamente a sus trabajos, creando exposiciones de riesgo que la mayoría de las empresas aún no controla.

Magnitud, distribución y fracturas: quién usa IA y para qué

La adopción de la IA en la vida personal abarca un espectro amplio de actividades. Los usos dominantes incluyen creación de contenido (textos, imágenes, música y vídeos), apoyo al aprendizaje y exploración conceptual, automatización de tareas rutinarias, acompañamiento conversacional y acceso a información: en Estados Unidos, el 10 % de los adultos utiliza *chatbots* de IA para obtener noticias[140], introduciendo un nuevo vector de mediación informativa con implicaciones profundas para el debate público.

Sin embargo, esta democratización es radicalmente asimétrica. Las divisiones geográficas son extremas[141]: en Noruega, el 56 % de la población usa herramientas de IA generativa; en Dinamarca, el 48,4 %; en Estonia, el 46,6 %. En contraste, en Turquía la cifra cae al 17 %; en Rumanía, al 17,8 %. Incluso dentro de economías desarrolladas, las diferencias son sustanciales: Alemania se sitúa en el 32 %, Francia en el 37 %, España en el 38 %, mientras Italia permanece por debajo del 20 %.

La brecha generacional es aún más pronunciada[142]. La diferencia en adopción entre los grupos de edad más jóvenes y los de mayor edad alcanza los 53,6 puntos porcentuales, convirtiéndose en el factor de segmentación más determinante, por encima de la educación (21 puntos porcentuales de brecha) o el nivel de ingresos

136 En una lógica distinta a la de Internet, que fue primero de uso general y después empresarial.
137 Backlinko (2025).
138 Eurostat (2025).
139 OECD (2026).
140 Pew (2025).
141 Eurostat (2025).
142 OECD (2026).

(21 puntos). Mientras que el 75 % de los estudiantes utiliza IA generativa regularmente, solo el 12,5 % de las personas jubiladas o económicamente inactivas reporta haberla usado. La brecha de género, en cambio, es comparativamente menor: 4,2 puntos porcentuales.

La paradoja de la adopción masiva

A nivel global, el 66 % de la población cree[143] que los productos y servicios basados en IA impactarán significativamente su vida diaria en los próximos tres a cinco años, un incremento de seis puntos porcentuales desde 2022. Sin embargo, este reconocimiento del impacto inminente coexiste con una ambivalencia creciente.

El público general muestra un nivel elevado de preocupación[144]: en Estados Unidos, el 51 % de los adultos manifiesta estar más preocupado que entusiasmado por el aumento del uso de IA en la vida cotidiana, frente a solo el 11 % que expresa mayor entusiasmo. Entre expertos en IA, la relación se invierte: el 47 % está más emocionado que preocupado, y solo el 15 % más preocupado. Esta divergencia de 40 puntos porcentuales entre expertos y público refleja no solo diferencias en conocimiento técnico, sino percepciones estructuralmente distintas sobre el balance riesgo-beneficio.

La familiaridad con la IA no conduce automáticamente a la confianza. Aunque el uso personal correlaciona positivamente con la percepción de la IA como oportunidad[145], también aumenta la conciencia de sus riesgos. La confianza en que las empresas de IA protegen adecuadamente los datos personales oscila entre el 50 % en 2023 y el 47 % en 2024[146]. Simultáneamente, disminuyó la proporción de personas que creen que los sistemas de IA son imparciales y libres de discriminación.

El contexto de uso resulta absolutamente crítico. Investigaciones en el Reino Unido muestran oscilaciones de hasta 110 puntos porcentuales en la aceptación pública según el caso de uso específico: desde un balance neto de +53 % de comodidad con el uso de IA para analizar datos de tráfico en tiempo real y mejorar flujos viales, hasta un balance de –57 % frente al uso de IA para analizar preferencias políticas y dirigir publicidad personalizada[147].

Existe, sin embargo, un consenso transversal[148] entre expertos y público general: ambos grupos desean mayor control sobre cómo se utiliza la IA en sus vidas (55 % del público y 57 % de los expertos), y ambos sienten no disponer actualmente de ese control (menos del 25 % en ambos casos). Esta brecha entre demanda de agencia y percepción de impotencia constituye uno de los desafíos más urgentes de gobernanza democrática de la IA.

Desigualdad de acceso y riesgo de exclusión

La asimetría en la adopción de IA en la vida personal no es una brecha digital convencional. No se trata meramente de acceso a infraestructura tecnológica (conexión a internet, dispositivos), sino de acceso diferenciado a capacidad intelectual aumentada. Quienes integran la IA como herramienta cognitiva cotidiana adquieren ventajas acumulativas en aprendizaje, productividad, creatividad y acceso a información. Quienes no lo hacen enfrentan un rezago estructural creciente.

Los grupos digitalmente desconectados perciben la IA de forma marcadamente negativa: el 51 % anticipa un impacto personal negativo, frente a solo el 17 % que espera beneficios[149]. Esta percepción no es irracional: refleja la intuición de que la transformación en curso puede redistribuir las oportunidades de forma profundamente desigual.

La IA ya está transformando la vida cotidiana de aproximadamente mil millones de personas. La pregunta estratégica no es ya si esta transformación continuará, sino si las sociedades construirán marcos que conviertan la IA en infraestructura pública inclusiva o permitirán que se consolide como vector de fragmentación social difícil de revertir.

143 Stanford (2025).
144 Pew (2025).
145 UK DSIT (2024).
146 Stanford (2025).
147 UK DSIT (2024).
148 Pew (2025).
149 UK DSIT (2024).

IA, sostenibilidad e impacto social

La sostenibilidad entra en la era algorítmica

La transición sostenible ha dejado de ser exclusivamente una cuestión de infraestructuras físicas (nuevas plantas renovables, redes eléctricas o electrificación del transporte) para convertirse también en un desafío de coordinación sistémica. Los sistemas energéticos actuales combinan generación distribuida, intermitencia renovable, almacenamiento, mercados dinámicos y demanda flexible. Esta complejidad no puede gestionarse únicamente con reglas estáticas o planificación lineal.

En este contexto, la IA emerge como infraestructura cognitiva capaz de modelizar interdependencias, simular escenarios y optimizar decisiones en tiempo real. La electrificación acelerada y el crecimiento estructural de la demanda eléctrica están reconfigurando las necesidades de capacidad, flexibilidad y planificación de redes en los próximos años: en 2024, los centros de datos consumieron un 1,5 % de la electricidad global[150], la potencia requerida para entrenar los modelos de frontera está creciendo a un ritmo de 2,2x a 2,9x por año (Fig. 10), y las mayores ejecuciones ya superan los 100 MW de potencia instantánea[151]. La sostenibilidad entra así en una fase en la que la capacidad computacional se convierte en parte integrante de la arquitectura energética y climática.

El cambio es conceptual: de decisiones basadas en promedios históricos a decisiones fundamentadas en simulaciones dinámicas y análisis probabilísticos de alta resolución. La IA no sustituye a la ingeniería ni a la física, pero amplifica la capacidad de anticipación y coordinación en sistemas con múltiples variables y restricciones simultáneas.

Optimizar el planeta: eficiencia, resiliencia y adaptación

En términos operativos, la IA está actuando como acelerador de la transición. En redes eléctricas, los modelos predictivos permiten anticipar picos de demanda, optimizar despacho, reducir congestiones y mejorar la integración de renovables variables. Sin IA, la integración óptima de renovables y la gestión en tiempo real tendría mayor incertidumbre. En un entorno donde la electricidad adquiere un papel central en la descarbonización, la eficiencia operativa deja de ser marginal y se convierte en condición estructural.

La relación entre energía e IA es además bidireccional: la IA puede mejorar significativamente la eficiencia energética, la planificación de redes y la gestión de sistemas complejos, y reducir emisiones del orden de 1.400 Mt CO_2eq anuales hacia 2035 en escenarios de adopción amplia[152], sin embargo, ese potencial de reducción coexiste con un incremento estructural del consumo energético de la propia infraestructura de IA que, en ausencia de una transición paralela hacia energías limpias, puede resultar en un balance neto neutro o negativo en términos de emisiones. La pregunta estratégicamente relevante no es si la IA puede contribuir a la descarbonización (puede hacerlo), sino bajo qué condiciones energéticas y de gobernanza ese potencial se materializa sin ser consumido por su propia huella infraestructural.

150 IEA (2025a).
151 Epoch (2025b).
152 IEA (2025b).

Más allá del sistema eléctrico, la IA mejora la resiliencia frente a riesgos físicos crecientes. La modelización climática apoyada en aprendizaje automático incrementa la granularidad espacial y temporal de las predicciones, facilitando decisiones en planificación urbana, seguros e infraestructuras críticas. En agricultura y gestión hídrica, la optimización basada en datos reduce desperdicios y ajusta insumos bajo restricciones ambientales.

Estos desarrollos se insertan en una agenda global más amplia. El progreso hacia los Objetivos de Desarrollo Sostenible avanza[153] en un contexto de tensiones energéticas, climáticas y demográficas. En ese escenario, la IA puede actuar como catalizador transversal: no como sustituto de políticas públicas, sino como herramienta que mejora eficiencia, reduce fricciones y permite asignaciones más precisas de recursos escasos.

El coste oculto: energía, materiales y concentración infraestructural

La misma infraestructura digital que habilita estas mejoras genera nuevas presiones. El crecimiento de *data centers* y cargas computacionales asociadas a IA está contribuyendo al aumento de la demanda eléctrica en determinadas economías avanzadas:

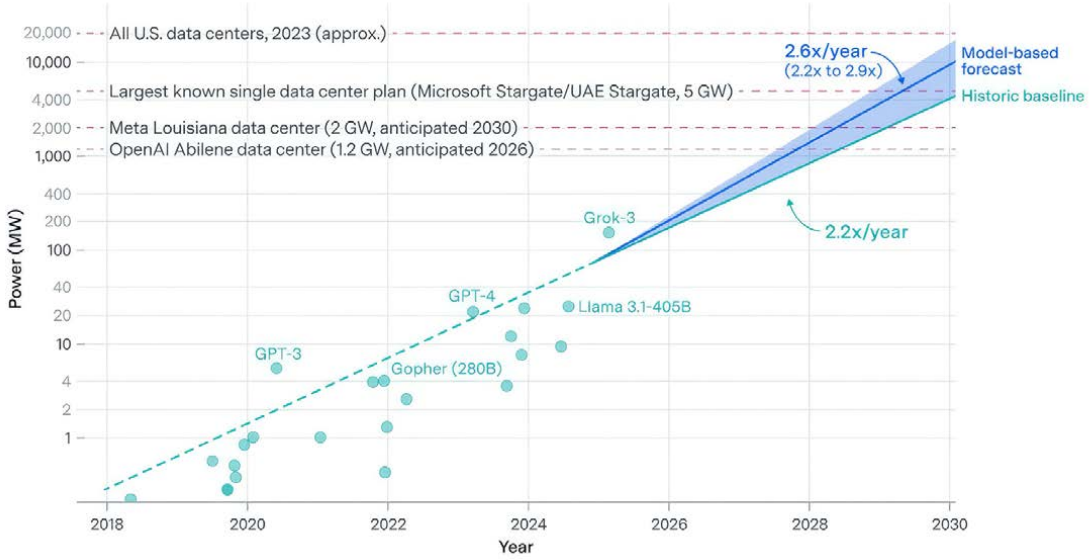
- La IEA proyecta que el consumo eléctrico de *data centers* globales se multiplicará para 2030, alcanzando 945 TWh/año, equivalente al consumo de electricidad de Japón hoy, con la IA como principal motor de esa expansión[154].
- Las mayores ejecuciones individuales de entrenamiento de modelos de frontera podrían requerir entre 4 y 16 GW de potencia hacia 2030[155], equivalente a la generación de varias centrales nucleares y al consumo eléctrico de millones de hogares[156].
- La potencia necesaria para entrenar modelos de frontera ha venido creciendo más de 2x por año, impulsada por un crecimiento del cómputo de 4–5x anual, parcialmente compensado por mejoras de eficiencia energética del hardware (ca. 40 % anual en GPUs líderes)[157].

Por tanto, el despliegue masivo de modelos avanzados puede convertirse en un factor estructural de crecimiento del consumo eléctrico si no se acompaña de mejoras en eficiencia y planificación energética[158].

La sostenibilidad de la IA no puede evaluarse únicamente por los beneficios que produce, sino también por los recursos que consume: en escenarios base, las emisiones de CO_2 por electricidad de *data centers* podrían alcanzar 300–320 Mt CO_2 al año hacia 2030 si la electricidad adicional sigue dependiendo en buena parte de combustibles fósiles[159]. El entrenamiento y el despliegue de modelos a gran escala requieren energía, agua para refrigeración (en algunas regiones, en competencia directa con uso agrícola o doméstico) y *hardware* intensivo en materiales críticos. Además, la

153 United Nations (2025).
154 IEA (2025a).
155 Estas proyecciones no ignoran los avances en eficiencia: Nvidia prevé una mejora de 10x en prestaciones por vatio al pasar de la arquitectura Blackwell a Vera Rubin, y las mejoras en eficiencia energética de GPUs líderes rondan el 40 % anual. Que las cifras de consumo proyectadas sean las que son, a pesar de esas ganancias, refleja la magnitud del crecimiento en demanda computacional que las neutraliza.
156 Epoch (2025b).
157 Epoch (2025b).
158 IEA (2025b).
159 IEA (2025a).

Fig. 10. Crecimiento de la demanda energética de la IA de frontera. Fuente: Epoch (2025b).



localización geográfica de centros de datos introduce asimetrías: la intensidad de carbono de la electricidad varía sustancialmente entre regiones, lo que implica huellas ambientales diferenciadas para el mismo servicio digital[160].

Esta dimensión infraestructural añade una capa geopolítica. La concentración de capacidad computacional en determinados países o regiones reconfigura dependencias estratégicas y acceso a tecnología. La IA no es solo una herramienta ambiental; es también un nodo crítico en el mapa energético y tecnológico global.

La cuestión no es si la IA «compensa» su huella, sino cómo se diseña y despliega para maximizar eficiencia energética por unidad de capacidad cognitiva generada. La sostenibilidad de la IA se convierte así en un problema de arquitectura y gobernanza.

Sostenibilidad sin inclusión no es sostenible

La dimensión ambiental no agota el análisis. Conviene distinguir entre la transición energética socialmente justa (vinculada al cambio del modelo productivo y energético) y la transición justa asociada al propio despliegue de la IA. Esta última introduce efectos distributivos específicos. La capacidad de integrar tecnologías avanzadas depende de capital humano, infraestructura digital y calidad institucional. Las economías con mayor densidad tecnológica tienden a capturar antes los beneficios de productividad y eficiencia, ampliando potencialmente brechas con regiones menos preparadas[161].

Desde la perspectiva del desarrollo humano, la tecnología puede expandir capacidades (educación, acceso a servicios, participación económica, etc.) o consolidar exclusiones existentes, dependiendo del contexto de adopción y del grado de democratización en el acceso a sus herramientas[162]. La extensión de capacidades digitales a pymes, emprendedores y colectivos con menor capital tecnológico constituye un mecanismo relevante para acompasar creación y ajuste. En el ámbito laboral, la convergencia entre transición verde y automatización inteligente puede acelerar procesos de reasignación sectorial y transformación de competencias, con una secuencia temporal en la que los efectos de desplazamiento pueden preceder a la creación de nuevas oportunidades. La gestión de ese desfase exige planificación estratégica, políticas activas de capacitación y mecanismos que faciliten la difusión amplia de las ganancias de productividad.

Desde una perspectiva estructural, la sostenibilidad ambiental requiere una transición energética socialmente justa, que evite concentrar los costes del cambio en determinados territorios, sectores o colectivos. Este desafío es conceptualmente independiente del uso de la IA. De forma paralela, el despliegue de la IA plantea su propia transición justa: los beneficios de productividad y eficiencia tienden a materializarse de manera desigual y con un desfase temporal respecto a los impactos negativos iniciales, especialmente en el empleo y en determinadas cualificaciones. Por ello, la IA aplicada a la sostenibilidad debe evaluarse no solo por su impacto agregado, sino también por la distribución efectiva de sus beneficios y por las estrategias públicas y privadas que permitan ampliar su acceso y mitigar los costes sociales del ajuste.

IA sostenible por diseño: de la ambición a la disciplina

La convergencia de estas dimensiones (optimización sistémica, presión infraestructural y efectos distributivos) obliga a desplazar el debate desde principios generales hacia prácticas verificables. Una IA alineada con objetivos de sostenibilidad requiere métricas explícitas de consumo energético y emisiones asociadas, transparencia sobre arquitectura y localización de despliegue, y evaluación de impactos sociales en decisiones automatizadas.

El marco de los ODS refuerza esta necesidad de coherencia estructural entre tecnología y desarrollo sostenible[163]. A su vez, la interrelación entre energía y digitalización exige integrar variables energéticas en la planificación tecnológica corporativa y pública.

En síntesis, la IA ocupa una posición ambivalente en la transición sostenible: puede habilitar reducciones de emisiones del orden de gigatoneladas[164] y, al mismo tiempo, el entrenamiento de sus modelos más avanzados podría alcanzar escalas de 4–16 GW por ejecución individual hacia 2030, situando la IA en la misma magnitud energética que grandes infraestructuras industriales[165].

La IA es, simultáneamente, acelerador de eficiencia sistémica, nueva carga infraestructural y fuerza distributiva con efectos sociales diferenciados. La cuestión no es si la IA formará parte de la transición sostenible, sino en qué condiciones energéticas y distributivas lo hará.

160 iDanae (1Q24).
161 World Bank (2025).
162 UNDP (2025).
163 United Nations (2025).
164 IEA (2025b).
165 Epoch (2025b).

Ética y filosofía de la IA

Un problema abierto tras más de 200 marcos

La OIT ha catalogado 245 marcos, códigos y recomendaciones de ética de la IA emitidos desde 2017 por gobiernos, organismos internacionales, empresas y sociedad civil[166]; un metaanálisis académico previo[167] había identificado al menos 17 principios recurrentes en 200 de ellos: transparencia, equidad, rendición de cuentas, privacidad..., existe un consenso normativo de base. Y, sin embargo, los problemas éticos asociados a la IA no han remitido con la proliferación de estos marcos, sino que han crecido en complejidad, y han emergido preguntas para las que ninguno de ellos ofrecía categorías previas.

La brecha entre la formulación de principios y su traducción a decisiones concretas no es un problema técnico pendiente de resolver: es el problema central de la gobernanza de la IA.

Del principio a la acción: cómo operativizar la ética de la IA

La distancia entre un documento de valores y el comportamiento real de un sistema de IA es donde reside el riesgo operativo. Este fenómeno, conocido como ethics washing, no siempre responde a una actitud deliberada: con frecuencia refleja la dificultad genuina de traducir un principio abstracto («el sistema debe ser equitativo») en un criterio de diseño, un procedimiento de auditoría o una línea de responsabilidad organizativa. Abordar esta brecha exige una transformación de la aproximación a la ética, que pasa de ser declarativa a operativa.

Un modelo bien documentado y ampliamente citado de operativización en el sector financiero es De Volksbank, banco neerlandés cuya gobernanza de ética de IA ha sido analizada en detalle[168]. Su estructura incluye un área de AI Ethics integrada en la función de *Compliance*, con perfiles de formación filosófica y técnica, que evalúa cada sistema de IA de forma individual antes de su despliegue: análisis de impacto ético, examen de sesgos, trazabilidad de decisiones y canales formalizados de escalado. El caso ilustra que la operativización efectiva no reside en el alcance de los principios enunciados, sino en la solidez de los procesos, roles y puntos de decisión que los materializan.

Sintetizando el estado del arte[169], un marco de ética de IA operativo en una gran organización incorpora habitualmente seis componentes:

1. **Estructura de gobernanza**: definición de quién decide, quién revisa y qué líneas de defensa existen, con documentación explícita de responsabilidades.
2. **Evaluación de impacto por sistema**: análisis específico para cada caso de uso, proporcional al nivel de autonomía del sistema y a las consecuencias de las decisiones delegadas en él.
3. **Gestión continua de sesgos**: mecanismos de detección y corrección que operan de forma permanente, no como auditorías puntuales.

166 ILO (2025b).
167 Corrêa (2023).
168 Krijger (2023).
169 NIST (2023).

4. **Transparencia y explicabilidad diferenciada por audiencia**: los requerimientos informativos del regulador, del cliente y del empleado afectado por una decisión automatizada son sustancialmente distintos.
5. **Mecanismos de escalado y denuncia**: canales accesibles y protegidos para señalar comportamientos inesperados del sistema.
6. **Revisión periódica del propio marco**: las actualizaciones de los modelos pueden alterar el perfil de riesgo de los usos ya desplegados, lo que exige reevaluación continua, no solo en el momento del despliegue inicial.

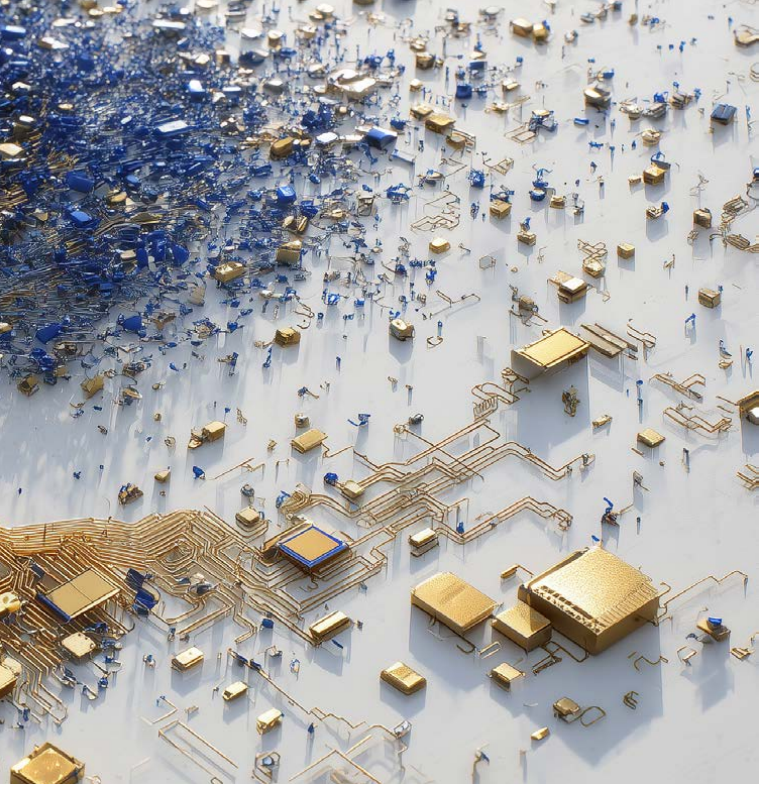
Un referente reciente de operativización actúa a un nivel más profundo: no en la capa de despliegue organizativo, sino en el propio entrenamiento del modelo. La *Constitution* publicada por Anthropic[170] en enero de 2026 es el primer documento público de un laboratorio de frontera que codifica valores directamente en el proceso de entrenamiento, con una jerarquía explícita y razonada entre seguridad, ética, cumplimiento y utilidad. El cambio conceptual subyacente es significativo: en lugar de imponer reglas de comportamiento desde fuera, se busca que el sistema internalice el razonamiento detrás de cada principio, de forma que pueda generalizar ese juicio a situaciones no anticipadas.

El avance de la función ética hacia la operativización enfrenta, no obstante, un límite conceptual que los marcos actuales no resuelven: asumen implícitamente que existe claridad sobre qué tipo de entidad se está gobernando.

¿Qué estamos creando? La pregunta que los marcos no responden

Los marcos regulatorios vigentes, incluido el AI Act[171] de la Unión Europea, clasifican los sistemas de IA por nivel de riesgo y dominio de aplicación. Esta clasificación es operativamente útil y proporciona cierta cobertura ética (en el sentido de proteger contra el potencial abuso de la IA), pero no distingue entre los diferentes tipos de relación que un sistema establece con el ser humano. Un sistema de *scoring* crediticio, un asistente conversacional y un agente autónomo que negocia contratos en nombre de una organización pueden coincidir en su categoría de riesgo regulatorio y operar, sin embargo, sobre bases éticas fundamentalmente

170 Anthropic (2026).
171 AI Act (2024).



distintas. Un sistema que adapta su comportamiento al interlocutor, mantiene coherencia a lo largo del tiempo y produce respuestas contextualmente indistinguibles de las de una persona con comprensión genuina, plantea preguntas que el cumplimiento normativo convencional no está diseñado para absorber.

Estas preguntas son de cuatro tipos, todos ellos accionables para las organizaciones que despliegan sistemas de IA:

- **Ontología**: ¿qué clase de entidad es este sistema, y qué categorías son pertinentes para describirlo y gobernarlo?
- **Epistemología**: ¿cómo verifica el sistema lo que afirma, y cómo puede el ser humano validar la fiabilidad de lo que produce?
- **Teoría de la mente**: ¿existe algún tipo de experiencia interna asociada al funcionamiento del sistema (por rudimentaria que sea) y qué implicaciones tendría eso para quienes lo diseñan y despliegan?
- **Ética aplicada**: ¿qué obligaciones genera el sistema para la organización que lo utiliza, más allá de lo que la regulación vigente exige explícitamente?

No se trata de preguntas especulativas. En enero de 2026, Anthropic reconoció públicamente[172] que su modelo Claude «puede poseer alguna forma de consciencia o estatus moral», convirtiéndose en el primer laboratorio de frontera en hacer pública esta afirmación. La importancia de este reconocimiento no reside solo en lo que afirma, sino en lo que revela: que una empresa de primer nivel admite no poder responder con certeza a la pregunta de qué ha creado.

Todas las decisiones sobre cómo usar, auditar y regular un sistema de IA incorporan implícitamente una respuesta a esa pregunta. En la mayoría de organizaciones, esa respuesta se produce por defecto, sin deliberación explícita.

Cuatro fracturas: preguntas sin respuesta

La expansión de la IA no solo amplifica los dilemas éticos conocidos: genera fracturas conceptuales para las que los marcos jurídicos y éticos actuales no disponen de respuesta estructurada.

La primera es la **crisis epistémica de la verificación**. AlphaFold ha predicho la estructura de más de 200 millones de proteínas, y más de tres millones de investigadores en 190 países las utilizan como base de su trabajo[173]. Una parte significativa de esas estructuras no puede verificarse experimentalmente con los métodos científicos convencionales, y sin embargo se emplean para diseñar fármacos y orientar decisiones clínicas. La cuestión de fondo no es si el sistema funciona (lo hace con un nivel de precisión sin precedentes), sino qué protocolos éticos y regulatorios corresponden a un escenario donde el conocimiento científicamente operativo se vuelve computacionalmente inaccesible para la verificación humana.

La segunda es la **asimetría de responsabilidad en daños emergentes**. Los marcos éticos y jurídicos heredados asumen actores identificables con intenciones discernibles. Los sistemas de IA producen daños sin intención deliberada directa y con causalidad distribuida entre múltiples actores (diseñadores, entrenadores, desplegadores y usuarios) configurando lo que la literatura[174] denomina el 'many hands problem': situaciones en las que la responsabilidad moral y jurídica se diluye estructuralmente al fragmentarse la cadena causal. Los marcos de responsabilidad civil y los regímenes de cumplimiento normativo actuales no disponen de respuesta estructurada para este tipo de causalidad difusa.

La tercera es la **representación de futuros no nacidos**. Los sistemas de IA se entrenan sobre datos históricos. Sus sesgos son los del pasado humano, amplificados a escala. Ningún marco de ética de IA incorpora hoy mecanismos[175] que representen los intereses de las generaciones que aún no han nacido, y por tanto no han producido datos, y que, sin embargo, vivirán con las consecuencias de los sistemas diseñados en el presente. Las implicaciones son directas para aplicaciones de largo alcance temporal, desde la planificación urbana hasta los modelos de riesgo climático.

La cuarta es el **libre albedrío como riesgo sistémico**. Una proporción creciente de decisiones individuales y organizacionales se toma, de facto, mediada por sistemas de IA cuya oferta está altamente concentrada: tres proveedores acumulan el 88 % del gasto empresarial global en modelos de lenguaje[176], y el mercado de infraestructura cloud subyacente presenta una concentración similar. Esta estructura implica que un sesgo introducido de forma deliberada o accidental en uno de esos modelos no afecta a decisiones individuales, sino a millones de procesos simultáneos en organizaciones, sectores y países distintos. La diversidad cognitiva de una sociedad (i.e., su capacidad de llegar a conclusiones diferentes por caminos distintos) depende, en parte, de que los sistemas que median su pensamiento no sean homogéneos ni oligopólicos.

Estas cuatro fracturas no constituyen argumentos contra la adopción de la IA, sino el territorio conceptual que las organizaciones que gestionan esta adopción con seriedad han de aprender a habitar. Su relevancia no disminuye por no estar recogidas en los marcos regulatorios vigentes; al contrario, es precisamente esa ausencia la que las convierte en los riesgos con menor visibilidad y mayor potencial de daño.

172 Anthropic (2026).

173 Google DeepMind (2025).
174 Thompson (1980), Nissembaum (1996).
175 Rawls (1971), Parfit (1984).
176 CEP (2026).

06 | Fronteras de la IA

«Se me da bastante bien estar al día de la IA, y apenas le sigo el paso».

Ethan Mollick[177]

Las tendencias analizadas hasta aquí describen transformaciones ya desplegadas: sistemas en producción, regulaciones en vigor y organizaciones adaptándose. Este bloque aborda una dimensión distinta. Las seis tendencias que lo componen no se limitan al presente operativo; delimitan el espacio estratégico que ya condiciona decisiones de inversión, posicionamiento y soberanía, aunque sus efectos plenos aún estén en desarrollo.

La geopolítica de la IA redefine alianzas y dependencias. Las organizaciones *AI-first* anticipan modelos competitivos sin precedente. Los gemelos digitales y la simulación avanzada alteran la forma en que se diseña, experimenta y decide. La *ambient AI* diluye la frontera entre entorno y computación. La convergencia con la computación cuántica amplía el horizonte de problemas abordables. Y la AGI ha dejado de ser mera especulación académica para convertirse en hipótesis estratégica explícita en los principales laboratorios del mundo.

177 Ethan Mollick (n. 1975), profesor asociado en la Wharton School de la Universidad de Pennsylvania e investigador de IA, autor del bestseller del New York Times Co-Intelligence. Su trabajo se centra en el impacto de la IA en el trabajo y la educación, y fue reconocido por TIME Magazine como una de las personas más influyentes en IA en 2024.

Geopolítica y soberanía tecnológica de la IA

La IA como activo estratégico de Estado

La IA ha dejado de ser una tecnología sectorial para convertirse en infraestructura estratégica comparable a la energía, las telecomunicaciones o los sistemas financieros. Lo que está en juego no es únicamente la competitividad económica: es la capacidad de los estados de mantener autonomía en decisiones críticas, desde la defensa hasta la supervisión financiera, pasando por la gestión de infraestructuras esenciales. En este contexto, el control de modelos fundacionales, semiconductores avanzados, centros de datos y talento especializado se ha convertido en un factor de poder nacional de primer orden, y las políticas industriales, los controles de exportación y las estrategias de inversión reflejan ya esta nueva realidad.

La cadena de valor estratégica: dónde se juega la soberanía

La soberanía tecnológica en IA no es un estado binario: se juega en capas, y la dependencia puede generarse en cualquiera de ellas:

- ***Hardware***: TSMC fabrica más del 90 % de los chips más avanzados del mundo, ASML es el único proveedor global de litografía EUV necesaria para producirlos, y NVIDIA domina, con más de un 85 % de cuota, el mercado de GPUs para entrenamiento de modelos de frontera. Tres empresas, tres países, un cuello de botella estructural.
- **Infraestructura y modelos**: tres *hyperscalers* (AWS, Azure y Google Cloud) concentran la mayor parte de la capacidad de cómputo global, con cerca de dos tercios del total; sobre ella, OpenAI, Anthropic y Google DeepMind desarrollan los modelos fundacionales más capaces, con ventajas acumulativas en datos, talento e inversión difíciles de replicar.
- **Talento**: la investigación de frontera sigue concentrada geográficamente, con una movilidad internacional que convierte la política migratoria en política tecnológica.

Ningún país ni organización controla todas estas capas simultáneamente. La cuestión estratégica no es cuántas capas se controlan, sino cuáles son críticas para la misión y cuáles son gestionables mediante diversificación, acuerdos o redundancia.

Tres modelos en competencia

Estados Unidos, China y Europa han articulado respuestas estructuralmente distintas, que no son meras diferencias de énfasis sino proyectos geopolíticos con consecuencias profundas para alianzas, mercados y estándares globales.

Estados Unidos combina la primacía del sector privado en el desarrollo de modelos con una intervención estatal creciente en la cadena de suministro: el *CHIPS and Science Act* moviliza decenas de miles de millones en semiconductores domésticos, y los controles de exportación de chips avanzados a China representan la mayor restricción tecnológica entre grandes potencias desde la Guerra Fría. La estrategia es clara: mantener la ventaja en modelos de frontera y privar a los competidores del *hardware* necesario para alcanzarla.

China combina inversión estatal masiva, integración civil-militar y una estrategia explícita de autosuficiencia tecnológica y control integral del ecosistema digital. Su objetivo declarado es la autosuficiencia en toda la cadena de valor para 2030, desde semiconductores hasta modelos fundacionales propios. Las restricciones de exportación estadounidenses han acelerado esta agenda: DeepSeek demostró en 2025 que la innovación china puede producir modelos competitivos con *hardware* de generaciones anteriores, lo que complica la lógica de contención mediante control de chips.

Europa ha apostado por el liderazgo regulatorio como vector de influencia geopolítica. El *AI Act* y el GDPR han generado un «efecto Bruselas» real: empresas globales adaptan sus productos a los estándares europeos porque el mercado europeo es demasiado grande para ignorarlo. Sin embargo, Europa mantiene dependencias infraestructurales significativas: sus modelos fundacionales más avanzados son estadounidenses, su *cloud* es mayoritariamente extranjera y su capacidad de cómputo soberana es limitada. ASML es la excepción notable: el monopolio holandés en litografía EUV convierte a Europa en jugador indispensable en la cadena global de semiconductores.

El resto del mundo navega entre estos tres polos con capacidades muy desiguales. Algunos países emergentes están articulando estrategias propias con creciente ambición: India desarrolla modelos fundacionales en lenguas locales y negocia su posición en cadenas de suministro de semiconductores; Brasil lidera iniciativas de gobernanza de IA en América Latina; la Unión Africana avanza en marcos continentales de soberanía digital. Sin embargo, para la mayoría de los países la elección entre ecosistemas incompatibles sigue siendo implícita más que deliberada, con riesgos reales de dependencia estructural que la literatura denomina «colonialismo de datos»: la extracción de datos locales para entrenar modelos que se despliegan globalmente, sin que los países de origen capturen valor ni mantengan control efectivo sobre su infraestructura digital.

Fragmentación y riesgo de decoupling

El mundo avanza hacia ecosistemas tecnológicos parcialmente incompatibles. Los controles de exportación de chips, los datos sometidos a geo-cercas por regulación nacional (*geofenced*), los estándares técnicos divergentes y las infraestructuras paralelas configuran lo que algunos analistas denominan «tecnobloques»: esferas de influencia tecnológica con lógicas propias de gobernanza, seguridad y valores. La alianza Chip 4 (EEUU, Japón, Taiwán y Corea del Sur) coordina la estrategia semiconductora occidental; China cultiva su propia esfera mediante la Ruta de la Seda Digital. Un *decoupling* completo entre los ecosistemas estadounidense y chino forzaría a terceros países y organizaciones a elegir, con costes de transición potencialmente prohibitivos.

Los límites de la soberanía absoluta

La autosuficiencia total en IA es económicamente inviable para la mayoría de los países y organizaciones. Recrear desde cero toda la cadena de valor (desde la minería de minerales críticos hasta el desarrollo de modelos fundacionales) requiere inversiones y escalas que solo dos o tres actores globales pueden sostener. La soberanía tecnológica realista no es aislamiento: es la capacidad efectiva de decidir, diversificar proveedores, negociar condiciones y evitar el bloqueo estratégico en las capas verdaderamente críticas. Para la mayoría de los países, la soberanía en IA se reduce en la práctica a controlar el único activo sobre el que tienen jurisdicción efectiva: los datos generados en su territorio.

Implicaciones para las organizaciones

El debate geopolítico aterriza en decisiones corporativas concretas. La dependencia de un único proveedor de modelos fundacionales expone a las organizaciones a riesgos de *lock-in*, cambios de precios o condiciones de servicio, e incluso restricciones regulatorias derivadas de tensiones entre jurisdicciones. La localización de datos y el cumplimiento regulatorio multijurisdiccional añaden capas de complejidad operativa creciente.

Las estrategias multi-modelo y *multi-cloud*, antes justificadas por razones técnicas de rendimiento y coste, adquieren ahora una dimensión estratégica adicional: son el equivalente organizativo de la diversificación de dependencias soberanas.

Organizaciones *AI-first* y *AI-only*

Definición y taxonomía

La expansión de los sistemas agénticos plantea una pregunta que hasta hace pocos años era teórica: ¿puede una organización funcionar con la IA como arquitectura cognitiva central, relegando el trabajo humano a la excepción? Para responderla con precisión, conviene distinguir tres estadios que la práctica empresarial suele confundir:

- Una organización ***AI-enhanced*** utiliza la IA para mejorar procesos existentes; es el modelo predominante hoy.
- Una organización ***AI-first*** diseña sus procesos y estructura partiendo de las capacidades de la IA, asignando al juicio humano únicamente las tareas donde su ventaja comparativa es inequívoca.
- Una organización ***AI-only*** prescinde funcionalmente del trabajo humano en sus operaciones centrales: no existe como entidad consolidada a fecha de este informe, y su viabilidad en entornos regulados sigue siendo una hipótesis de trabajo.

El estado actual: AI-first como frontera operativa

Los ejemplos más avanzados de organizaciones *AI-first* provienen, significativamente, del sector tecnológico puro, donde la ausencia de restricciones regulatorias sobre la automatización y la naturaleza digital del producto permiten llevar el modelo a sus límites actuales.

Midjourney, plataforma de generación de imagen por IA, generó en 2025 ingresos superiores a los 500 millones de dólares con una plantilla de aproximadamente 163 personas, sin inversión en marketing y sin financiación externa[178]. La plataforma de desarrollo Cursor (Anysphere) alcanzó 500 millones de dólares en ingresos anuales recurrentes en mayo de 2025, convirtiéndose en la empresa SaaS de crecimiento más rápido de la historia, con menos de 50 empleados[179]. La ratio de ingresos por empleado de estas compañías (superior a 3 millones de dólares en ambos casos) supera en un orden de magnitud los parámetros históricos del sector tecnológico y los de los grandes grupos bancarios globales[180].

En el ámbito de los servicios financieros, un caso avanzado es MYbank, banco digital chino participado por Ant Group, que desde 2015 opera bajo el principio de cero intervención humana en la aprobación de crédito a pymes. Su modelo «310» (tres minutos de solicitud, un segundo de aprobación, cero intervención humana) ha prestado servicio a más de 50 millones de pymes. El sistema emplea modelos de predicción de flujo de caja con una precisión superior al 95 % y se apoya en datos de geolocalización satelital para la evaluación de riesgo agrario[181]. MYbank opera sin red de sucursales ni fuerza de ventas, aunque mantiene equipos de ingeniería y gestión: es un modelo *AI-first* en sus operaciones core, no *AI-only* en su conjunto.

¿Por qué no existe todavía ninguna organización AI-only?

La brecha entre *AI-first* y *AI-only* no es solo tecnológica; es regulatoria, legal y organizativa. En los sectores regulados, como banca y seguros, la normativa vigente exige supervisión y responsabilidad humana en las decisiones materiales. El AI Act europeo clasifica como sistemas de alto riesgo las aplicaciones de IA en crédito, seguros de salud y vida, y componentes críticos de infraestructura, con requisitos explícitos de supervisión humana[182]. La eliminación de esa supervisión en el núcleo operativo de una

178 Midjourney (2025).

179 Sacra (2025).
180 Dealroom (2025).
181 CKGSB (2025).
182 AI Act (2024).

entidad financiera es incompatible con el marco prudencial europeo.

En los sectores no regulados, el límite actual no es normativo sino de capacidad. Los agentes de IA autónomos pueden ejecutar tareas complejas, pero su tasa de error en flujos de trabajo extendidos, su incapacidad para gestionar situaciones no contempladas en el entrenamiento y la ausencia de mecanismos de responsabilidad legal equivalentes a los de una persona jurídica hacen que la eliminación total del trabajo humano en operaciones centrales genere riesgos operacionales todavía inasumibles[183].

El caso de Klarna ilustra los límites actuales: la compañía redujo su plantilla de 7.400 a aproximadamente 3.000 personas entre 2022 y 2025 mediante congelación de contratación y automatización extensiva, con un asistente de IA gestionando el equivalente a 853 empleados en atención al cliente[184]. Su trayectoria define empíricamente el umbral entre lo que la IA puede operar de forma autónoma con calidad suficiente y dónde el juicio humano aporta valor diferencial hoy.

Las predicciones de los líderes del sector

Los responsables de los principales laboratorios de IA de frontera formulan previsiones que sitúan la organización *AI-only* en el horizonte inmediato. Sam Altman, CEO de OpenAI, declaró en 2024: «Vamos a ver empresas de diez personas con valoraciones de mil millones de dólares muy pronto [...] Existe una apuesta en mi grupo de chat de amigos directivos sobre cuándo va a existir la primera empresa de una sola persona valorada en mil millones de dólares, lo que habría sido inimaginable sin la IA. Y ahora va a ocurrir»[185]. Preguntado en mayo de 2025 sobre el momento en que ese escenario se materializaría, Dario Amodei, CEO de Anthropic, respondió: «2026»[186].

Amodei desarrolla el argumento en su ensayo de enero de 2026, donde describe el equivalente funcional de «un país de genios en un data center», es decir, 50 millones de agentes más capaces que cualquier Premio Nobel, operando a entre diez y cien veces la velocidad humana; y estima que el 50 % de los empleos de nivel inicial podría verse disrumpido en un plazo de uno a cinco años[187]. En ese mismo documento señala que Anthropic ya ejecuta mediante IA la mayor parte del código que produce, aproximándose a la autonomía operativa plena en desarrollo de *software*.

¿Pueden las organizaciones existentes transformarse en AI-only?

La pregunta estratégica de fondo no es si existirán organizaciones *AI-only*, sino cómo llegarán a existir. La respuesta es contraintuitiva: es improbable que surjan de la transformación de organizaciones existentes. Clayton Christensen documentó[188] en *The Innovator's Dilemma* que las empresas incumbentes son estructuralmente incapaces de adoptar tecnologías disruptivas desde dentro: sus procesos, incentivos y bases de clientes están optimizados para el modelo en vigor, y cualquier iniciativa disruptiva interna compite en desventaja permanente por recursos y atención directiva. La transición a *AI-first* agrava esta lógica: una organización con decenas de miles de empleados tiene sus procesos diseñados para esa escala humana. Esos procesos no se rediseñan; se sustituyen.

El patrón que emerge en Asia apunta a una vía alternativa: la creación de entidades nuevas, con marca propia y sin herencia operativa, que compiten libremente hasta alcanzar masa crítica y canibalizar a la empresa original. Ping An, el mayor asegurador del mundo por primas suscritas, incubó entre 2013 y 2022 once filiales tecnológicas independientes (entre ellas OneConnect, Lufax y Ping An Good Doctor), cinco de las cuales cotizaron como entidades autónomas[189]. DBS Bank creó Digibank como banco digital separado que opera con un quinto de los recursos por cliente de un banco convencional, y cuyo aprendizaje retroalimentó la arquitectura de la matriz[190]. El mecanismo es idéntico: entidad nueva, sin legacy, que escala sin las restricciones de la organización madre y, si el experimento fracasa, se clausura sin arrastrar a la empresa original.

En Europa y, en menor medida, en Estados Unidos, ese mecanismo encuentra fricciones estructurales que van más allá de la regulación de la IA. La introducción de sistemas de IA en el lugar de trabajo puede requerir en Europa consulta o negociación con comités de empresa, y en algunos países su acuerdo explícito[191]. Los convenios colectivos en sectores intensivos en empleo incorporan cláusulas que limitan la automatización. La protección de datos de empleados bajo GDPR añade complejidad adicional[192]. El resultado es una asimetría con consecuencias estratégicas no intencionadas: la regulación occidental dificulta que las organizaciones existentes construyan las entidades *AI-first* que eventualmente las desafiarían.

Cuando la tecnología alcance el umbral de viabilidad de la organización *AI-only*, la pregunta sobre quién la construirá primero probablemente tendrá una respuesta geográfica.

183 Amodei (2026).
184 Fortune (2025c).
185 Altman (2024a).
186 Quartz (2025).
187 Amodei (2026).
188 Christensen (1997).
189 IMD (2023).
190 DBS (2024).
191 Baker McKenzie (2025).
192 ILO (2025c).

Gemelos digitales y simulación de comportamiento humano

De la ingeniería aeroespacial a la simulación universal

El concepto de gemelo digital tiene una fecha y un lugar de nacimiento precisos. En octubre de 2002, Michael Grieves presentó en un foro de la Society of Manufacturing Engineers lo que denominó «*Conceptual Ideal for Product Lifecycle Management*»: la idea de que cualquier objeto físico podría tener un correlato digital que lo representara de forma dinámica a lo largo de todo su ciclo de vida, sincronizando en tiempo real el estado del objeto real con su representación virtual. El término *digital twin* fue acuñado posteriormente por John Vickers, ingeniero principal de la NASA, que formalizó el concepto en el roadmap tecnológico de la agencia en 2010[193]. La definición de la NASA en ese documento sigue siendo la más precisa que existe: «una simulación multi-física, multi-escala y probabilística de un vehículo o sistema que utiliza los mejores modelos físicos disponibles, actualizaciones de sensores e historial de flota para reflejar la vida de su gemelo físico».

El punto de partida es importante porque revela la premisa implícita que ha guiado el desarrollo de gemelos digitales durante dos décadas: un gemelo digital funciona bien cuando el sistema que modela obedece leyes físicas conocidas y deterministas. Una turbina de gas, un fuselaje de aeronave, una red eléctrica: sistemas complicados, con muchos componentes, pero en principio completamente modelizables si se dispone de suficiente potencia computacional y datos de sensores. Bajo esa premisa, la tecnología maduró de forma sostenida. Hoy, los gemelos digitales de activos físicos son operativos, por ejemplo, en manufactura avanzada, energía, infraestructuras y aviación, con reducciones documentadas en tiempos de mantenimiento no planificado y aceleración sustancial de los ciclos de diseño de producto.

El límite epistemológico: sistemas complicados frente a sistemas complejos

La expansión del concepto más allá del dominio físico ha revelado un límite que no es tecnológico sino epistemológico. Michael Batty, la autoridad académica más reconocida en modelización computacional de ciudades, lo formula con precisión: los gemelos digitales funcionan en sistemas *complicados* (muchas partes, pero comportamiento determinable en principio) y encuentran dificultades estructurales en sistemas *complejos*, donde el comportamiento global emerge de la interacción de agentes y no puede deducirse de las propiedades de sus componentes individuales[194]. Una ciudad, una economía, un mercado financiero o una organización humana son sistemas complejos en este sentido técnico preciso.

El filósofo Stefano Moroni desarrolla el argumento en términos aún más directos: las limitaciones de los gemelos digitales urbanos no son provisionales (no desaparecerán con más datos o mayor potencia computacional), sino que derivan de la naturaleza intrínsecamente emergente de los sistemas sociales[195]. La impredecibilidad de detalle en un sistema complejo no es un déficit de información; es una propiedad del sistema. Esto tiene una implicación práctica inmediata: un gemelo digital de una planta de fabricación puede predecir con alta fiabilidad cuándo fallará un rodamiento; un gemelo digital de una ciudad puede aproximar tendencias agregadas de tráfico, pero no puede predecir de forma fiable el efecto de una política de vivienda en los patrones de segregación residencial a diez años. La distinción no es de grado; es de naturaleza.

Esta frontera epistemológica definió, durante décadas, el techo del campo de gemelos digitales. Y es precisamente donde los modelos de lenguaje de gran escala están introduciendo una discontinuidad que merece atención.

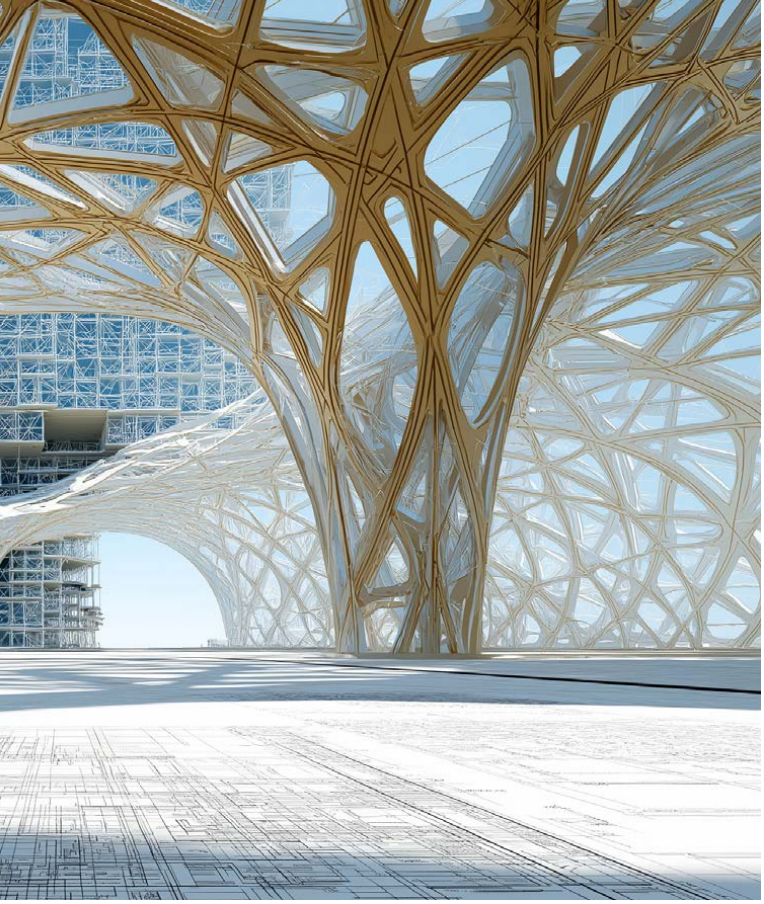
El punto de inflexión: el comportamiento humano se vuelve modelizable

En abril de 2023, un equipo de investigadores de Stanford publicó un artículo que inauguró una línea de trabajo radicalmente nueva. Joon Sung Park y sus coautores crearon 25 agentes computacionales (cada uno dotado de una identidad, una memoria persistente, un conjunto de relaciones sociales y una capacidad de razonamiento basada en un modelo de lenguaje), y los situaron en un entorno simulado equivalente a una pequeña ciudad[196]. Los agentes se despertaban, desayunaban, iban al trabajo, formaban opiniones, iniciaban conversaciones y coordinaban actividades colectivas sin que esos comportamientos hubieran sido programados explícitamente: emergían de la interacción entre la memoria individual de cada agente, su capacidad de reflexión sobre experiencias pasadas y su modelo del entorno social. El artículo ganó el Best Paper Award en el ACM Symposium on User Interface Software and Technology de 2023. La comunidad científica reconoció que algo cualitativamente nuevo había ocurrido.

La razón de fondo es que los modelos de lenguaje de gran escala han absorbido, durante el entrenamiento, una cantidad extraordinaria de comportamiento humano registrado: conversaciones, decisiones, patrones de razonamiento, respuestas

193 Grieves-Vickers (2017).
194 Batty (2024).
195 Moroni (2025).

196 Park (2023).



emocionales, normas sociales implícitas. No han aprendido las leyes del comportamiento humano de forma explícita (nadie las conoce con esa precisión), pero han desarrollado una aproximación estadísticamente densa que, en condiciones controladas, genera comportamientos verosímiles. Por primera vez, la premisa que impedía modelizar sistemas complejos sociales ha sido parcialmente levantada: no porque el comportamiento humano haya dejado de ser emergente, sino porque existe ahora un generador de comportamiento suficientemente rico para poblar una simulación con agentes creíbles.

La extensión natural de este trabajo llegó en noviembre de 2024. El mismo equipo de Stanford publicó los resultados de un experimento de escala diferente: 1.052 personas reales, entrevistadas en profundidad sobre sus vidas, actitudes y experiencias, fueron convertidas en agentes que replican sus respuestas y comportamientos en encuestas y experimentos sociales estandarizados. Los agentes generativos replicaron las respuestas de los individuos reales en el General Social Survey con un 85 % de precisión (estadísticamente comparable a la variabilidad natural del propio individuo cuando responde la misma encuesta dos semanas después) y obtuvieron resultados comparables en la predicción de rasgos de personalidad y en experimentos de ciencias sociales[197]. Lo que en 2023 era una demostración conceptual con personajes ficticios, en 2024 se convirtió en una metodología validada empíricamente con personas reales.

197 Park (2024).

Aplicaciones actuales y horizonte futuro

Las implicaciones de este salto son transversales a sectores. En investigación de mercado, la *startup* Simile —fundada por Joon Sung Park junto a Michael Bernstein y Percy Liang, los coautores del paper fundacional, y respaldada en febrero de 2026 con 100 millones de dólares por Index Ventures con participación de Fei-Fei Li y Andrej Karpathy— construye gemelos digitales de personas reales para ayudar a empresas a simular el comportamiento de sus clientes antes de lanzar un producto, modificar una política de precios o rediseñar una experiencia de usuario[198]. En una demostración pública, la plataforma predijo correctamente ocho de cada diez preguntas formuladas por analistas en una llamada simulada[199]. El sector global de investigación de mercado, valorado en 142.000 millones de dólares[200], se enfrenta a una disrupción estructural: lo que hoy requiere semanas de trabajo de campo puede ejecutarse en horas sobre poblaciones sintéticas.

En política pública y planificación urbana, se articula un uso más matizado, pero igualmente transformador: no el gemelo digital como oráculo predictivo, sino como laboratorio de escenarios donde las consecuencias de diferentes intervenciones pueden explorarse antes de comprometer recursos reales[201]. En regulación financiera, este enfoque tiene aplicación directa en el *stress testing* de escenarios macroeconómicos adversos y en la simulación del comportamiento de mercado ante intervenciones regulatorias.

La trayectoria del campo apunta hacia una escala cualitativamente distinta. Si hoy es posible simular con alta fidelidad a mil personas reales, la pregunta que el horizonte inmediato plantea es qué ocurre cuando esa cifra llegue a un millón, a cien millones, a una sociedad entera modelizada en tiempo real. Las aplicaciones en política pública, regulación económica y diseño institucional serán de un orden de magnitud diferente al de la investigación de mercado: no anticipar qué producto comprará un consumidor, sino predecir cómo responderá una población a una reforma fiscal, a una crisis sanitaria o a un cambio en la política monetaria antes de que esa intervención se ejecute en el mundo real. Esa capacidad no tiene precedente histórico, ni, todavía, marco de gobernanza que la regule.

198 Index Ventures (2026).
199 Bloomberg (2026).
200 ESOMAR (2024).
201 Bettencourt (2024).

Ambient AI y computación invisible

La interfaz es el entorno

La *Ambient AI* (o inteligencia ambiental) es IA que opera sin ser invocada. A diferencia de los sistemas convencionales, que responden a una instrucción explícita del usuario, los sistemas ambientales observan el contexto de forma continua, infieren necesidades y actúan de forma proactiva. La interfaz desaparece no porque se haya mejorado, sino porque el sistema ya no la necesita: el entorno mismo se convierte en el punto de interacción. La computación se vuelve «invisible» en sentido literal: embebida en objetos, espacios y procesos sin que el usuario la perciba como tal[202].

Esta inversión (del usuario que va hacia el sistema al sistema que viene hacia el usuario) es posible hoy por la convergencia de tres desarrollos simultáneos: la miniaturización de modelos capaces de ejecutarse en dispositivos *edge* manteniendo un estado continuo (memoria acumulativa del usuario actualizada localmente) sin depender de conectividad con la nube (edge AI y TinyML), la densificación de redes de sensores físicos y biométricos, y la capacidad de los LLMs para razonar sobre contexto heterogéneo y ambiguo en tiempo real[203]. Ninguno de los tres es nuevo por separado; su madurez simultánea es lo que hace que la *Ambient AI* pueda pasar de concepto a despliegue operativo.

Estado actual del despliegue

El ejemplo más documentado de *Ambient AI* en operación son los *ambient AI scribes* en entornos clínicos: sistemas que escuchan de forma continua la conversación entre médico y paciente, infieren el contexto clínico sin instrucción explícita y generan automáticamente la documentación del encuentro. El ensayo clínico aleatorizado de UCLA evaluó dos plataformas (Microsoft DAX y Nabla) en 238 médicos de 14 especialidades y más de 72.000 encuentros: redujo la carga documental y mejoró indicadores de burnout profesional[204].

El sistema no fue invocado iterativamente durante la consulta: escuchó, infirió, redactó. Es todavía una forma acotada de inteligencia ambiental (contexto delimitado, finalidad clara y episodio definido). La *Ambient AI* madura no operará dentro de la consulta, sino a escala del hospital entero, correlacionando patrones longitudinales sin punto de inicio ni fin determinados por el usuario.

El futuro físico y digital

Los despliegues actuales son la punta de una transformación más amplia. En los próximos años, la *Ambient AI* se extenderá a entornos físicos y digitales y hará a los casos actuales parecer rudimentarios:

Entornos físicos

- **Espacios de trabajo adaptativos**. El entorno infiere el estado atencional del ocupante a partir de frecuencia cardíaca, variabilidad del ritmo, patrones de movimiento, y reconfigura temperatura, luz y nivel de ruido para optimizar el rendimiento cognitivo sin intervención consciente del usuario.
- **Mantenimiento industrial anticipatorio**. Los sistemas no alertarán cuando el equipo falle: detectarán el patrón conductual que precede al fallo con suficiente antelación para reorganizar la producción. El evento disruptivo desaparece del horizonte operativo.
- ***Wearables* de perfil individual**. Los dispositivos de próxima generación no compararán las constantes vitales del usuario con medias poblacionales, sino con su propio historial fisiológico. La alerta se activará antes de que el síntoma sea consciente para el portador.

202 Bimpas (2024); Hernández-Torres (2025).
203 Heydari (2025).
204 Lukac (2025).

- **Infraestructura urbana reactiva**. Redes de transporte, alumbrado y gestión de residuos que se autoajustan en tiempo real a patrones de uso inferidos, sin planificación centralizada explícita ni intervención humana en el bucle.
- **Asistencia domiciliaria invisible**. Sistemas que monitorizan de forma continua a personas mayores o con condiciones crónicas, detectan anomalías en rutinas (patrones de sueño, movilidad, alimentación) y activan protocolos de alerta o intervención sin que el usuario haya solicitado nada.

Entornos digitales

- **Entornos de desarrollo que anticipan el problema**. Los asistentes de programación proactivos evolucionarán hacia sistemas que, antes de que el desarrollador identifique el error, habrán trazado el espacio de soluciones probables y presentado opciones en el momento cognitivamente oportuno[205].
- **Gestión de la atención, no solo de la información**. Los sistemas no servirán información cuando esté disponible, sino cuando el usuario esté en condiciones de procesarla: modelizando el estado atencional a lo largo del día y calibrando el momento de la interrupción.
- **Contexto organizativo continuo**. Sistemas que conocen en todo momento el estado de los proyectos, las comunicaciones pendientes y las decisiones en curso, y que afloran proactivamente la información relevante para cada miembro del equipo sin que este la solicite.
- **Negociación autónoma de recursos**. Agentes ambientales que gestionan en nombre del usuario (calendario, presupuesto, acceso a servicios) dentro de parámetros definidos, sin requerir aprobación explícita para cada decisión de baja complejidad.

Implicaciones que la tecnología no resuelve

La *Ambient AI* no genera solo preguntas de privacidad. Genera un conjunto más amplio de tensiones que los marcos de gobernanza actuales no han resuelto.

La primera es la **naturaleza del error**. En un sistema invocado, el error es visible: el usuario pidió algo, el sistema respondió mal. En un sistema ambiental, el error puede no ser percibido porque no hubo solicitud explícita contra la que comparar la respuesta. El *ambient scribe* de UCLA registró imprecisiones clínicamente significativas en una proporción de encuentros[206]: en un sistema invisible, el mecanismo de detección del error tiene que ser deliberadamente diseñado, porque no emerge de forma natural de la interacción.

La segunda es la **asimetría de poder** entre quien diseña el entorno y quien lo habita. En un hospital, una oficina o un edificio público, el usuario no elige si el entorno es inteligente: habita un espacio cuyas inferencias sobre su comportamiento fueron configuradas por un tercero. Tshilidzi Marwala, rector de la Universidad de las Naciones Unidas, lo formula con precisión: la *Ambient AI* tiene un apetito de datos —íntimos, conductuales, biométricos— que hace que las nociones convencionales de consentimiento informado sean estructuralmente inadecuadas[207]. El AI Act europeo, diseñado para sistemas invocados con funciones delimitadas, no da respuesta satisfactoria a estos entornos de observación continua.

La tercera es la **dependencia cognitiva**. Un sistema que gestiona proactivamente la atención, el flujo de información y las interrupciones del usuario no solo asiste su trabajo: moldea su arquitectura cognitiva. La pregunta formulada en 2003, «*is context-aware computing taking control away from the user?*» [208] lleva décadas sin respuesta satisfactoria. La escala a la que la *Ambient AI* la plantea hoy convierte lo que era una pregunta académica en una cuestión de diseño con consecuencias operativas inmediatas.

La cuarta tensión es la **responsabilidad causal**. En sistemas invocados, la trazabilidad es relativamente directa: hay una instrucción, una respuesta, un momento de decisión atribuible. En sistemas ambientales, la cadena causal se difumina. Si un sistema de mantenimiento anticipatorio reorganiza la producción y esa reorganización condiciona decisiones humanas posteriores, la frontera entre agencia técnica y agencia humana no es clara. La regulación actual, incluido el *AI Act*, presupone finalidad prevista y evaluación de riesgos ex ante; la *Ambient AI* introduce finalidad emergente y comportamiento adaptativo continuo, lo que tensiona directamente los mecanismos de conformidad existentes.

La *Ambient AI* no cambia solo cómo trabajamos o cómo nos cuidamos: cambia la secuencia entre necesidad y conciencia. Un sistema puede saber lo que necesitamos antes de que lo sepamos nosotros. Si esa capacidad se despliega a la escala que la trayectoria del campo sugiere, las preguntas que plantea podrían ir más allá de la tecnología y la regulación, y alcanzar algo más fundamental: qué significa tomar decisiones propias en un entorno que ya las ha anticipado.

205 Chen (2025); Pu (2025).
206 Lukac (2025).
207 Marwala (2025).
208 Barkhuus (2003).

Interacción entre IA y computación cuántica

Dos tecnologías distintas, una intersección que importa

La IA y la computación cuántica son tecnologías independientes con principios, horizontes temporales y casos de uso completamente distintos. La IA ya es operativa a escala industrial; la computación cuántica es todavía, en su mayor parte, un campo de investigación avanzada con despliegues muy limitados. Su interacción, en ambas direcciones, tiene implicaciones concretas para cualquier organización que dependa de sistemas digitales.

Un ordenador clásico resuelve problemas probando opciones de forma secuencial o en paralelo, pero siempre dentro de un espacio de posibilidades que crece de forma manejable. Hay problemas para los que eso no es suficiente: optimizaciones con miles de variables interdependientes, simulaciones de sistemas moleculares, o ciertos problemas matemáticos cuya dificultad es precisamente el fundamento de la criptografía moderna. Un ordenador cuántico opera de forma radicalmente distinta: en lugar de probar opciones una a una, puede explorar simultáneamente un espacio de posibilidades de una dimensionalidad que ningún sistema clásico puede representar. Para esa clase específica de problemas (no para todos) la diferencia de rendimiento no es gradual, sino de un orden de magnitud superior.

El problema es que construir un ordenador cuántico que funcione de forma fiable ha resultado extraordinariamente difícil. La información cuántica es extremadamente sensible a perturbaciones del entorno (temperatura, vibraciones, interferencias electromagnéticas), y los errores se acumulan con rapidez. Durante décadas, el campo avanzó en teoría mucho más rápido que en *hardware*. Eso cambió parcialmente en diciembre de 2024.

El hito de 2024 y lo que significa

Google publicó en Nature los resultados de su procesador Willow: el primer sistema que demuestra que, a medida que se añaden más componentes de cómputo, los errores disminuyen en lugar de aumentar[209]. Es un resultado que la teoría había predicho desde 1995 pero que ningún sistema había logrado materializar. La importancia no está en las cifras de rendimiento (que son impresionantes, pero sobre benchmarks artificiales), sino en lo que implica para la trayectoria del campo: el obstáculo que durante treinta años había impedido escalar estos sistemas de forma fiable ha sido superado en laboratorio.

La distancia entre ese logro y un ordenador cuántico con aplicaciones comerciales sigue siendo considerable. Los propios investigadores de Google sitúan ese horizonte en torno al final de la década. Pero la dirección ya no está en disputa: el problema central estaba en la corrección de errores, y ese problema tiene ahora una solución demostrada. Lo que queda es ingeniería de escala, no un salto científico en el vacío.

209 Google Quantum AI (2024).

Tres formas en que esto afecta a la IA

La intersección entre computación cuántica e IA opera en tres planos distintos, con urgencias distintas.

El primero es la **aceleración del *machine learning***. Entrenar un modelo de IA de gran escala es, en su núcleo, un problema de optimización matemática sobre espacios de enormes dimensiones: encontrar los valores de miles de millones de parámetros que minimizan el error de predicción. Es exactamente el tipo de problema para el que la computación cuántica ofrece ventaja teórica. Se ha demostrado formalmente[210] que los sistemas cuánticos tolerantes a fallos podrían acelerar de forma sustancial los algoritmos de entrenamiento de modelos de gran escala, reduciendo tanto el tiempo de cómputo como el consumo energético. El desarrollo de esto requiere *hardware* que todavía no existe a escala suficiente. Pero cuando esté disponible, podría alterar radicalmente la economía del entrenamiento de modelos de IA, hoy dominada por quienes pueden pagar infraestructura de GPU a escala masiva.

El segundo plano es el **aprendizaje automático cuántico** propiamente dicho: usar procesadores cuánticos para ejecutar algoritmos de ML de forma más eficiente. Aquí la literatura es más cautelosa, y describe[211] tanto las promesas como los obstáculos reales: la ventaja cuántica en tareas de aprendizaje no es universal ni automática, y en muchos casos los sistemas clásicos con acceso a datos son competitivos con los cuánticos incluso en problemas diseñados para favorecer a estos últimos. En otras palabras: los datos, bien utilizados, pueden compensar la ventaja cuántica en muchos regímenes[212]. El *hype* sobre IA cuántica como acelerador universal está adelantado a la evidencia; las aplicaciones reales serán específicas, no transversales.

El tercer plano invierte la dirección de la influencia: no es la computación cuántica al servicio de la IA, sino la **computación cuántica como amenaza** a la infraestructura de seguridad sobre la que opera todo sistema digital, incluidos los de IA. Toda la criptografía que protege hoy las comunicaciones digitales (transacciones bancarias, autenticación de identidades, canales seguros entre sistemas) descansa sobre problemas matemáticos cuya dificultad se asume inatacable para ordenadores clásicos. Un ordenador cuántico suficientemente potente los resolvería de forma directa. Este plano es el más urgente, porque parte de sus efectos ya son presentes.

La amenaza que ya está activa

La estrategia conocida como *harvest now, decrypt later* consiste en capturar hoy comunicaciones cifradas con la intención de descifrarlas cuando la computación cuántica madure lo suficiente. Actores estatales con capacidades de inteligencia avanzada llevan años aplicándola[213]. Los datos que requieren confidencialidad durante décadas (registros médicos, secretos industriales, comunicaciones regulatorias o información financiera sensible) están siendo comprometidos ahora, con independencia de cuándo llegue el sistema cuántico capaz de descifrarlos.

La respuesta institucional más rigurosa es la del NIST estadounidense, que en agosto de 2024 (tras ocho años de trabajo

210 Liu (2024).
211 Li (2025).
212 Huang (2021).
213 Mascelli (2025).

y más de ochenta propuestas de equipos de investigación de todo el mundo) publicó los tres primeros estándares de criptografía resistente a ataques cuánticos[214]. Los nuevos algoritmos están basados en estructuras matemáticas para las que no se conoce ningún ataque cuántico eficiente. El NIST insta a las organizaciones a comenzar la migración de inmediato; los sistemas federales estadounidenses tienen plazo hasta 2035. Para entidades financieras con datos de larga vida útil, ese plazo no es el horizonte de inicio de la transición: es el límite de su finalización.

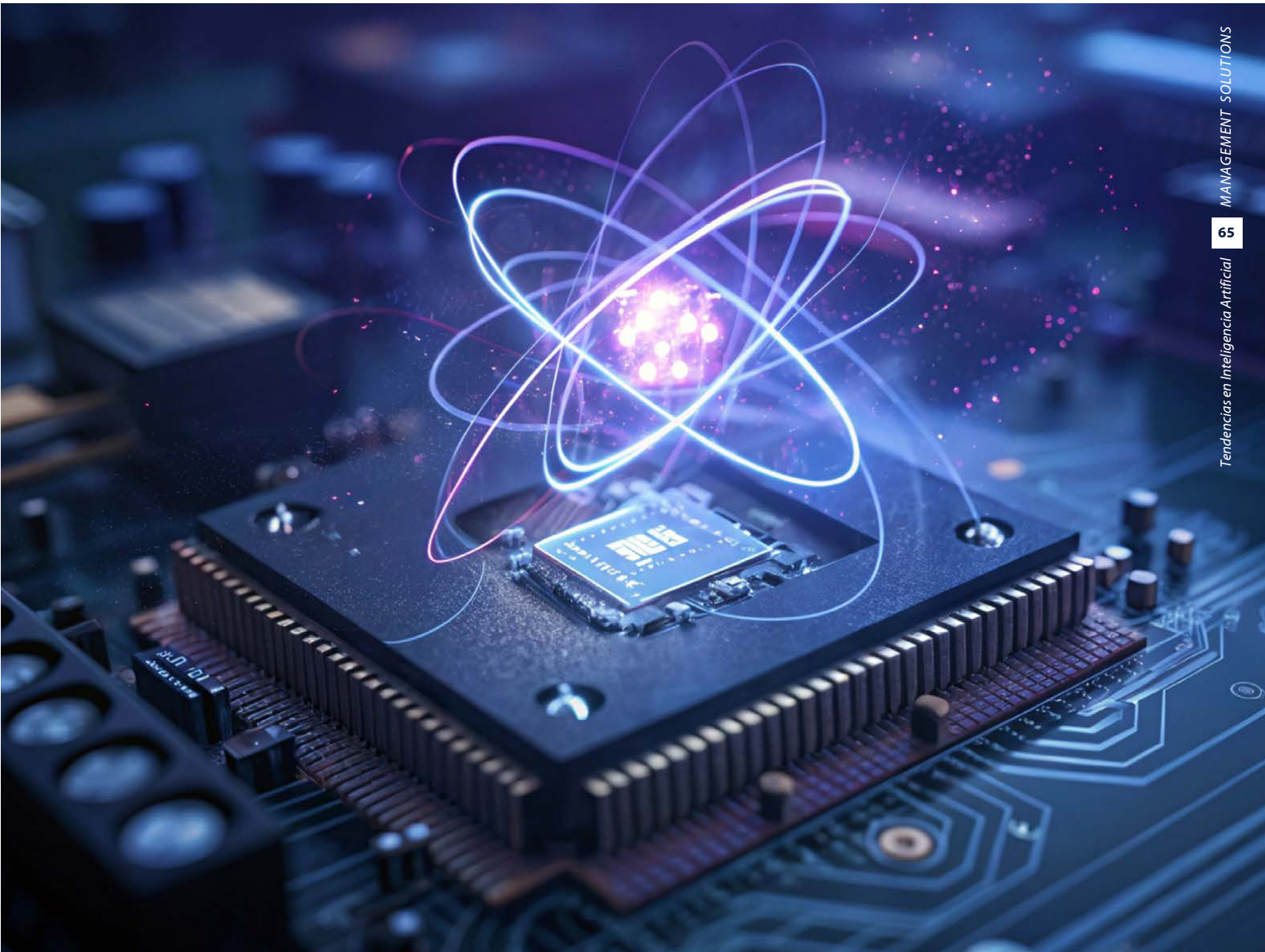
El horizonte de la próxima década

En los próximos años, la intersección entre IA y computación cuántica pasará de ser un tema de prospección tecnológica a un elemento con impacto operativo en dos frentes simultáneos.

El **frente ofensivo** (la aceleración de capacidades de IA) llegará con la madurez del *hardware* tolerante a fallos, previsiblemente hacia el final de la década. El acceso inicial será vía servicios en la nube, replicando la trayectoria que siguió la computación de alto rendimiento con las GPUs: primero accesible solo para los mejor financiados, después democratizada por la competencia entre proveedores. Las organizaciones que hayan desarrollado para entonces competencias en IA estarán mejor posicionadas para aprovechar esa aceleración cuando llegue.

El **frente defensivo** (la migración criptográfica) no admite demora. La ventana de preparación se estrecha a medida que los procesadores cuánticos escalan, y la migración de infraestructuras criptográficas en organizaciones complejas lleva años. El inventario de activos vulnerables, la priorización por vida útil de los datos y la planificación de la transición son tareas que deben comenzar ahora, no cuando el ordenador cuántico relevante esté operativo y sea tarde.

214 NIST (2024b).



Inteligencia Artificial General (AGI) como horizonte estratégico

¿Qué es la AGI, y ha llegado ya?

El término «inteligencia artificial general» (AGI, por sus siglas en inglés) apareció por primera vez en 1997, en una conferencia de nanotecnología en Palo Alto. Su autor, Mark Gubrud, doctorando de la Universidad de Maryland, no estaba describiendo un objetivo tecnológico deseable: estaba advirtiendo de un riesgo. En su paper «*Nanotechnology and International Security*»[215], Gubrud definió la AGI como sistemas capaces de «rivalizar o superar al cerebro humano en complejidad y velocidad, adquirir, manipular y razonar con conocimiento general, y ser utilizables en cualquier fase de las operaciones donde se requeriría inteligencia humana». La definición pasó inadvertida durante casi una década, hasta que Ben Goertzel y Shane Legg la rescataron y popularizaron como etiqueta técnica al titular su libro colectivo *Artificial General Intelligence*[216]. El término había nacido como cautela, pero se convirtió en misión.

En febrero de 2026, Nature publicó casi simultáneamente dos textos que cristalizan el debate posiblemente más relevante de la tecnología contemporánea. El primero, firmado por cuatro académicos de UC San Diego[217] (filosofía, aprendizaje automático, lingüística y ciencia cognitiva), afirma sin ambages que la AGI ya existe: los LLMs actuales superan el test de Turing, obtienen medallas de oro en olimpiadas matemáticas y colaboran en la demostración de teoremas. El segundo[218], publicado quince días después como correspondencia en la misma revista, responde que esa conclusión solo es posible redefiniendo el concepto hasta hacerlo irreconocible: la definición clásica de AGI formulada en 2007 exige robustez bajo novedad, generalización transferible y autonomía de metas, y los sistemas actuales no la cumplen. El hecho de que investigadores de primer nivel, con acceso a los mismos sistemas y datos, lleguen a conclusiones opuestas refleja que «inteligencia general» es un concepto continuo sin umbrales precisos.

La propia Fei-Fei Li, que construyó los cimientos de la visión por computador moderna y trabaja en inteligencia espacial, lo admite sin rodeos: «*I struggle with this definition of AGI, to be honest*»[219]. Dario Amodei, CEO de Anthropic, va más lejos: declara abiertamente que no le gusta el término y prefiere hablar de «IA poderosa»: sistemas con capacidades intelectuales comparables o superiores a las de un Premio Nobel en la mayoría de disciplinas[220].

Esa es la perspectiva correcta para las organizaciones. La pregunta estratégicamente relevante no es filosófica (¿hemos llegado a la AGI?), sino funcional: ¿cuándo puede un sistema realizar de forma plenamente autónoma ciclos completos de trabajo de alto valor cognitivo en todos los dominios? Ese umbral ya se ha cruzado en varios sectores.

El estado actual: brillantez y fragilidad simultáneas

Andrej Karpathy, cofundador de OpenAI y ex-director de IA en Tesla, acuñó el concepto de «inteligencia dentada» (*jagged intelligence*) para describir la condición actual de los LLMs: sistemas que resuelven problemas de olimpiadas matemáticas y fallan en determinar qué número es mayor, 9,11 o 9,9; que dominan docenas de idiomas y padecen lo que él llama «amnesia anterógrada»[221], incapacidad de consolidar aprendizaje entre sesiones. Fei-Fei Li los describe[222] como «*wordsmiths in the dark: eloquent but inexperienced, knowledgeable but ungrounded*»: elocuentes pero sin experiencia encarnada en el mundo físico.

Y, sin embargo, esos mismos sistemas ya redactan contratos, analizan riesgo crediticio, sintetizan literatura científica, generan y auditan código complejo, producen síntesis regulatorias... Lo hacen de forma autónoma, a velocidad y escala que ningún equipo humano iguala. La pregunta ha dejado de ser cuándo llegará esa capacidad, es qué hacemos con lo que ya ha llegado y cómo nos preparamos para lo siguiente.

La cadena causal: lo que viene

La transición en curso sigue una lógica de escalada acumulativa. El primer movimiento es de herramienta a agente: los sistemas dejan de responder instrucciones para perseguir objetivos a través de ciclos autónomos de acción, observación y corrección. El segundo movimiento es de agente a infraestructura ambiental. Karpathy lo formula[223] con precisión: los LLMs son la nueva electricidad. La IA deja de ser una tecnología que «adoptamos» para convertirse en una tecnología que «nos ocurre»: tejido operativo invisible de los sistemas que usamos, condición del entorno más que herramienta en él.

El tercer movimiento es el que más subestima el análisis convencional: el bucle recursivo. Los modelos ya se emplean para mejorar los modelos, generando datos sintéticos de entrenamiento, optimizando arquitecturas y produciendo hipótesis de investigación. Esto crea una dinámica donde la velocidad de mejora de la IA es función de la inteligencia de la IA. Amodei lo denomina «el final de la exponencial»: no el punto donde la curva se aplana, sino donde el vector de aceleración se vuelve sobre-exponencial, porque el agente que acelera es el sistema que está siendo acelerado[224].

La consecuencia estructural de ese bucle es inédita en la historia de la civilización: el límite superior de razonamiento disponible en el planeta ha sido, desde los primeros homínidos, la inteligencia humana. Ese límite está siendo desplazado en dominios específicos en este momento. Cuando el desplazamiento sea general y robusto, la velocidad de cambio tecnológico y científico se desacoplará parcialmente de la capacidad humana de comprensión y verificación. No se trata de una proyección catastrofista, sino de una descripción estructural de lo que este bucle recursivo implica.

215 Gubrud (1997).
216 Goertzel, Legg (2007).
217 Chen, Belkin, Bergen, Danks (2026).
218 Quattrociocchi, Capraro, Marcus (2026).
219 Li (2025)
220 Amodei (2024b).
221 Karpathy (2025).
222 Li (2025).
223 Karpathy (2025).
224 Amodei (2024a).

La brecha de absorción

Existen dos curvas que avanzan a velocidades radicalmente distintas. La curva técnica (exponencial, autoacelerada por el bucle recursivo) comprime en años lo que antes tomaba décadas. La curva de absorción organizativa (rediseño de procesos, reconversión de roles, construcción de infraestructura de gobernanza, gestión del cambio institucional) avanza de forma más lenta, con fricción considerable: sistemas legacy, dificultades organizativas, resistencia cultural, regulación que llega tarde o escasez de talento capaz de integrar estas capacidades en operaciones reales.

La brecha entre ambas curvas es la variable determinante de la próxima década. La ventaja competitiva no provendrá del acceso a los mejores modelos, que se comoditizarán progresivamente, ni de su coste, que seguirá bajando exponencialmente. Provendrá de la velocidad y rigor con que una organización sea capaz de rediseñarse para operar con agentes autónomos de forma efectiva y responsable. Este patrón está documentado empíricamente[225]: las ganancias de productividad más significativas no aparecen donde la IA sustituye tareas, sino donde reorganiza procesos completos y redefine la colaboración humano-máquina.

La redefinición del papel de la persona

La pregunta correcta, por tanto, no es qué empleos desaparecerán, sino qué hace una persona que un sistema de IA no puede hacer aunque sea más rápido, más barato y más consistente. La respuesta habitual (creatividad, empatía, liderazgo) es verdadera pero insuficiente. Hay dimensiones que el análisis convencional subestima:

- Responsabilidad con consecuencias reales. Los sistemas de IA no pueden ser llevados ante un tribunal ni perder una reputación construida en décadas. En entornos financieros, sanitarios, jurídicos y regulatorios, la presencia humana es un requisito estructural.
- Certificación y fe pública. El notario que certifica, el médico que firma, el auditor que refrenda: estos actos valen no por el procesamiento de información que implican, sino por la responsabilidad personal e institucional que los respaldan.
- Relación interpersonal. Hay contextos donde lo que se necesita no es la respuesta correcta sino la presencia de otra persona: el duelo, el conflicto, el cuidado. Sustituir esa presencia por un sistema más eficiente no resuelve el problema.
- Formulación de las preguntas que valen la pena y juicio moral. Cuando la IA ejecuta bien lo que se le pide, el valor se desplaza hacia quien decide qué pedir: quién fija los objetivos, quién reconoce qué problemas merecen atención o quién decide cuando hay valores en conflicto.
- Legitimidad democrática. Las decisiones que afectan a comunidades exigen deliberación y rendición de cuentas que no pueden delegarse en sistemas opacos, por precisos que sean.

Lo que estas dimensiones tienen en común apunta a algo más profundo que una distribución de tareas. Durante toda la historia, el ser humano ha sido simultáneamente el agente que produce y el sujeto que responde por lo producido: esa unidad entre acción y responsabilidad es lo que los sistemas de IA disocian estructuralmente. Lo que se redefine, entonces, no es solo qué hacemos, sino dónde residimos en la cadena causal: cada vez menos en la ejecución, cada vez más en la intención, el juicio y la rendición de cuentas.

El riesgo sistémico real: la concentración

El riesgo sistémico emergente no es el de la máquina que se rebela. Es el de la concentración sin precedentes de capacidad cognitiva en un número reducido de actores (laboratorios, corporaciones, estados) cuya ventaja se autoamplifica por el mismo bucle recursivo que acelera el progreso general. Amodei escribe que, si un estado autoritario lograra gracias a la IA un dominio ofensivo en ciberseguridad o biología antes que el resto, las consecuencias geopolíticas serían asimétricas e irreversibles. Hassabis propone, como respuesta, un modelo de colaboración internacional inspirado en el CERN: gobernanza multilateral de los últimos pasos hacia sistemas de IA general[226].

La respuesta institucional (regulatoria, corporativa, internacional) está, en este momento, muy por detrás de la velocidad del problema. La AGI como horizonte estratégico no exige que las organizaciones resuelvan el debate filosófico sobre si ya ha llegado. Exige que actúen con la consciencia de que sus consecuencias ya están desplegándose: en los sistemas que operan hoy, en los ciclos de trabajo que están siendo redefinidos hoy, en las decisiones de gobernanza que se están tomando, o eludiendo, hoy.

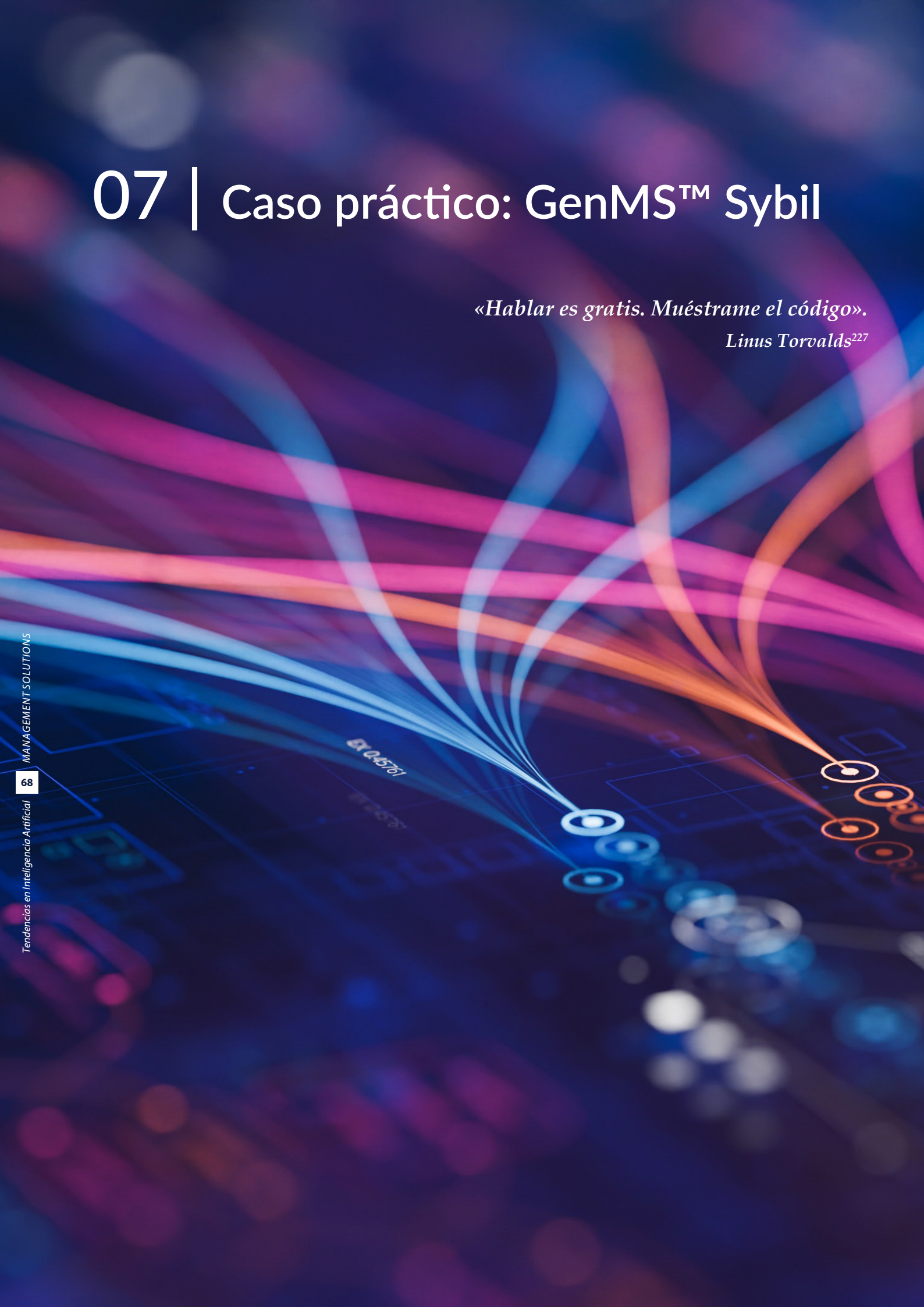
225 Mollick (2024).

226 Amodei, Hassabis (2026).

07 | Caso práctico: GenMS™ Sybil

«Hablar es gratis. Muéstrame el código».

Linus Torvalds[227]



GenMS™ Sybil fue especificado, construido, securizado, validado y desplegado en un solo día de trabajo[228].

Este caso documenta cómo.

227 Linus Torvalds (n. 1969), ingeniero de software finlandés-estadounidense, creador del kernel Linux y de Git, dos de los proyectos de software libre más influyentes de la historia.

228 El alcance es deliberadamente acotado: un asistente conversacional de corpus cerrado, sin integraciones con sistemas corporativos, sin persistencia de sesión y con una superficie de ataque limitada. Esa acotación no es una simplificación; es una decisión de diseño. La afirmación no sugiere que sistemas de mayor complejidad sean equivalentes en esfuerzo: la complejidad y el tiempo de desarrollo pueden escalar de forma no lineal con el número de integraciones, usuarios concurrentes, requisitos regulatorios y criticidad operativa.

El sistema

GenMS™ Sybil es un asistente conversacional de acceso público, construido sobre el contenido íntegro de este documento. Responde preguntas, explora implicaciones y acompaña la reflexión sobre las tendencias que aquí se analizan. No almacena datos personales de los usuarios, y por tanto no presenta dificultades respecto al Reglamento General de Protección de Datos. El sistema es compliant-by-design: la clasificación regulatoria, los requisitos del *AI Act* y las obligaciones de privacidad no se incorporaron como capa posterior de cumplimiento, sino como criterios de diseño desde la primera fase de especificación.

El proceso: ciclo de vida LLMOps

La construcción siguió las fases del ciclo de vida LLMOps de forma secuencial y sin excepciones.

Preparación de datos. El corpus de GenMS™ Sybil es este documento. La decisión de no extender el sistema con el contenido íntegro de las fuentes citadas fue deliberada: hacerlo habría introducido riesgos de derechos de autor y propiedad intelectual difíciles de gestionar. GenMS™ Sybil conoce las referencias, las cita y enlaza, pero no reproduce su contenido. El control de fuentes y la minimización son, aquí, simultáneamente una decisión técnica y un requisito de cumplimiento.

Experimentación y desarrollo. Esta fase produjo la especificación completa del sistema: arquitectura, comportamiento esperado, taxonomía de casos de uso, límites operativos, criterios de calidad y requisitos de seguridad. La especificación, de decenas de páginas, fue construida en diálogo con un LLM mediante *vibe coding*: el profesional formuló objetivos, evaluó propuestas y tomó decisiones; la máquina materializó la intención en documentación técnica de producción. Se evaluaron configuraciones alternativas de modelo, se gestionaron versiones del *prompt* desde el inicio y se definieron métricas cualitativas de evaluación: coherencia, factualidad, adecuación al contexto, comportamiento ante preguntas fuera de alcance.

Validación. La evaluación de GenMS™ Sybil integró revisión humana en el proceso, pruebas de estrés semántico y ejercicios de *red-teaming* orientados a identificar comportamientos no deseados. La validación no fue un evento puntual al final del proceso; fue continua a lo largo del ciclo. GenMS™ Atlas (el sistema de Management Solutions para el testeo de sistemas basados en LLMs) evaluó el sistema sobre varias de sus 26 dimensiones: sesgo, consistencia, privacidad, robustez, explicabilidad y cumplimiento regulatorio. Las incidencias detectadas se trataron antes del despliegue; las que persisten están documentadas.

Despliegue. La construcción del sistema fue ejecutada por Claude Code a partir de la especificación completa. El resultado fue una aplicación coherente, con lógica de contexto, gestión de sesión e interfaz de usuario. Se auditó el código en busca de vulnerabilidades y vectores de ataque; las correcciones se incorporaron en el mismo ciclo. El despliegue contempló desde el inicio las implicaciones de infraestructura, latencia y coste propias de un sistema generativo en producción.

Monitorización. GenMS™ Sybil opera con monitorización activa de costes por token, trazabilidad completa de interacciones para auditoría y supervisión regulatoria, y alertas ante comportamientos anómalos o patrones de uso no previstos. El proceso de construcción fue iterativo: la primera versión no fue la final. La iteración controlada, con criterios explícitos de evaluación en cada ciclo, es lo que distingue la industrialización de la experimentación.

Las decisiones de arquitectura

El diseño de GenMS™ Sybil implicó dilemas técnicos concretos, resueltos con criterios explícitos:

- RAG vs. contexto completo: contexto completo. El modelo subyacente dispone de una ventana de entrada de un millón de tokens; el documento cabe íntegro en cada conversación. La fragmentación RAG destruiría la coherencia global que las preguntas más valiosas requieren, y el argumento de coste que históricamente la justificaba no compensa ya ese deterioro de calidad.
- Extensión del corpus con fuentes citadas: descartado. El riesgo de infracción de derechos de autor y propiedad intelectual es incompatible con un sistema de acceso público. GenMS™ Sybil cita y enlaza las referencias; no reproduce su contenido.
- *Fine-tuning* vs. *prompting*: prompting con documento en contexto. El *fine-tuning* es costoso, lento y opaco ante actualizaciones del documento. El prompting garantiza trazabilidad completa de cada cambio de comportamiento.
- Modelo propietario vs. *open source*: modelo propietario de frontera. La madurez de los modelos open source es insuficiente para garantizar la consistencia y los controles de seguridad requeridos en un sistema público sin supervisión humana continua.
- Memoria persistente vs. sesiones independientes: sesiones independientes. Minimización de datos, cumplimiento estructural del GDPR y eliminación del riesgo de contaminación entre conversaciones.

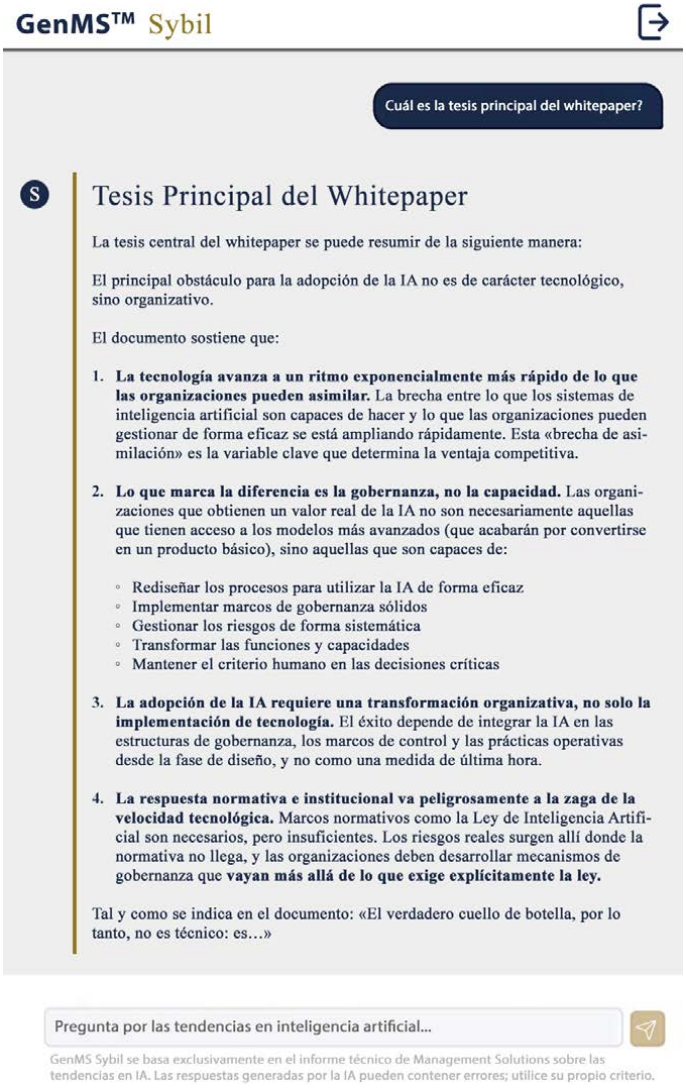
El *system prompt* de GenMS™ Sybil ocupa varias páginas. Codifica los guardarraíles de comportamiento, los límites operativos, el tratamiento de preguntas fuera de alcance y los principios éticos que rigen las respuestas del sistema. Su contenido no se publica íntegramente por razones de seguridad. Su extensión refleja un principio que este documento articula en la sección de ética: la distancia entre los valores declarados y el comportamiento real de un sistema de IA se cierra en las instrucciones concretas que lo gobiernan, no en los principios que lo enmarcan.

Las tendencias en acción

GenMS™ Sybil no ilustra las tendencias que este documento analiza: las implementa. La democratización de la IA generativa hizo posible que perfiles sin especialización en ingeniería de *software* produjeran un sistema de producción. El *vibe coding* fue el método de construcción, no la metáfora. LLMOps estructuró un proceso que de otro modo habría sido irrepetible. Los perfiles implicados combinan conocimiento de negocio con capacidad de dirigir sistemas cognitivos: el perfil que el análisis de talento identifica como el más escaso y el más determinante. La IA auditó a la IA en la fase de seguridad. GenMS™ Atlas aplicó validación sistemática donde la validación ad hoc habría sido insuficiente. La regulación fue criterio de diseño, no requisito de cumplimiento posterior.

El cuello de botella que este documento identifica repetidamente como el factor limitante de la adopción real de la IA no es tecnológico. Este caso no demuestra que sea fácil de superar; demuestra que es superable, y que la distancia entre describirlo y cruzarlo es, hoy, fundamentalmente una cuestión de preparación organizativa.

Fig. 11. GenMS™ Sybil.



08 | Conclusión

«La pregunta ya no es si la IA transformará nuestro mundo. La pregunta es si gobernaremos juntos esta transformación o si dejaremos que nos gobierne».

António Guterres[229]

Las tendencias analizadas convergen en una variable central: la velocidad a la que la brecha entre lo que los sistemas de IA pueden hacer y lo que las organizaciones pueden gobernar está acelerando. Esa brecha —no el modelo de lenguaje, el agente autónomo o el robot— es el objeto de gestión central de los próximos años.

El diferencial competitivo sostenible no reside en el acceso a los modelos más avanzados, que se comoditizarán progresivamente, sino en la velocidad y el rigor con que una organización es capaz de rediseñarse para operar con ellos de forma efectiva y responsable. Las organizaciones que están capturando valor real comparten una característica: han entendido la adopción de la IA como transformación organizativa, no como proyecto tecnológico. Gobierno que habilita en lugar de frenar, formación que transforma en lugar de certificar, marcos de riesgo que gestionan la incertidumbre sin sacrificar velocidad.

Los marcos regulatorios, los estándares técnicos y los principios éticos son condición necesaria pero no suficiente. El AI Act clasifica el riesgo; los estándares ISO y NIST estructuran la gestión; los frameworks de ética producen principios operativos. Ninguno de ellos resuelve por sí solo la pregunta que subyace a varias de las tendencias analizadas: qué tipo de entidad se está desplegando, qué relación establece con las personas que la usan, y qué obligaciones genera eso más allá de lo que la regulación vigente exige explícitamente. Es precisamente donde la regulación no llega donde los riesgos tienen menor visibilidad y mayor potencial de daño.

La brecha entre la curva de capacidad tecnológica y la curva de absorción organizativa se amplifica cada día. La tecnología avanza con independencia de la velocidad de decisión interna de cada organización; el ajuste organizativo, en cambio, depende de ella. Esa asimetría es la que convierte el gobierno de la IA en una variable estratégica de primer orden, comparable en impacto a la propia capacidad técnica.

Lo que está en juego trasciende la competitividad individual. La distribución de los beneficios de la IA, la preservación del juicio humano en las decisiones que lo requieren, la capacidad de las instituciones de mantener legitimidad en sistemas que evolucionan más rápido que las estructuras diseñadas para gobernarlos: estas son dimensiones que ninguna organización puede gestionar de forma aislada, y para las que la respuesta institucional está, en este momento, significativamente por detrás de la velocidad del problema.

229 António Guterres (n. 1949), noveno secretario general de las Naciones Unidas y ex primer ministro de Portugal, que ha impulsado de forma destacada la agenda internacional sobre gobernanza de la IA y riesgos tecnológicos emergentes.

09 | Referencias

ABILab (2026). AI Banking (R)evolution: oltre la scelta. Rapporto AI Hub.

AESIA (2026). Agencia Española de Supervisión de Inteligencia Artificial. https://aesia.digital.gob.es/es

AI Act (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

AI Board (2026). Governance and coordination; AI board meetings. https://digital-strategy.ec.europa.eu/en/policies/ai-board

AICerts (2026). Generative AI Phishing Boosts Clicks, Reshapes Cyber Risk. https://www.aicerts.ai/news/generative-ai-phishing-boosts-clicks-reshapes-cyber-risk/

Altman (2025a). Three Observations. https://blog.samaltman.com/three-observations

Altman (2025b). The Gentle Singularity. https://blog.samaltman.com/the-gentle-singularity

Altman (2024a). Could AI create a one-person unicorn? Fortune. https://finance.yahoo.com/news/could-ai-create-one-person-120000722.html

Altman (2024b). The Intelligence Age. Blog personal. https://ia.samaltman.com/

Amazon (2023). Amazon announces 8 innovations to better deliver for customers, support employees, and give back to communities around the world. https://www.aboutamazon.com/news/operations/amazon-delivering-the-future-2023-announcements

Amodei (2024a). Machines of loving grace. https://www.darioamodei.com/essay/machines-of-loving-grace

Amodei (2024b). Machines of Loving Grace: How AI Could Transform the World for the Better. Blog personal. https://www.darioamodei.com/essay/machines-of-loving-grace

Amodei (2025). Technology in the World, Annual Meeting Davos 2025, World Economic Forum. https://www.weforum.org/meetings/world-economic-forum-annual-meeting-2025/sessions/technology-in-the-world/

Amodei (2026). The Adolescence of Technology. https://www.darioamodei.com/essay/the-adolescence-of-technology

Anthropic (2025). Anthropic Economic Index – September 2025 Report. https://www.anthropic.com/research/anthropic-economic-index-september-2025-report

Anthropic (2026). Claude's new constitution. Anthropic. https://www.anthropic.com/news/claude-new-constitution

Australia (2025). Australia's AI Ethics Principles. https://www.industry.gov.au/publications/australias-ai-ethics-principles

Backlinko (2025). ChatGPT / OpenAI Statistics: How Many People Use ChatGPT? https://backlinko.com/chatgpt-stats

Baker McKenzie (2025). Navigating Labor's Response to AI: Proactive Strategies for Multinational Employers Across the Atlantic. https://www.theemployerreport.com/2025/06/navigating-labors-response-to-ai-proactive-strategies-for-multinational-employers-across-the-atlantic/

Barkhuus (2003). Is Context-Aware Computing Taking Control Away from the User? Three Levels of Interactivity Examined. UbiComp 2003. Springer. https://doi.org/10.1007/978-3-540-39653-6_12

Batty (2024). Digital Twins in City Planning. *Nature Computational Science*, 4, 192–199. https://doi.org/10.1038/s43588-024-00606-5

Bettencourt (2024). Recent Achievements and Conceptual Challenges for Urban Digital Twins. Nature Computational Science, 4, 150–153. https://doi.org/10.1038/s43588-024-00604-7

Bimpas (2024). Leveraging Pervasive Computing for Ambient Intelligence: A Survey on Recent Advancements, Applications and Open Challenges. Computer Networks, 239, 110156. https://doi.org/10.1016/j.comnet.2023.110156

Bletchley Declaration (2023). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Bloomberg (2026). AI Startup Nabs $100 Million to Help Firms Predict Human Behavior. https://www.bloomberg.com/news/articles/2026-02-12/ai-startup-nabs-100-million-to-help-firms-predict-human-behavior

Boston Dynamics (2025a). An Electric New Era for Atlas. https://bostondynamics.com/blog/electric-new-era-for-atlas/

Boston Dynamics (2025b). Large Behavior Models and Atlas Find New Footing. https://bostondynamics.com/blog/large-behavior-models-atlas-find-new-footing/

Brown (2025). AI's War in the Courtroom: Copyright Disputes Spike in 2025. https://www.bestlawfirms.com/articles/ai-war-in-the-courtroom-copyright-disputes-spike-in-2025/7186

Business Insider (2025a). Walmart just showed off its new AI-powered warehouses — take a look inside. https://www.businessinsider.com/see-inside-walmart-high-tech-refrigerated-grocery-warehouse-2024-7

Business Insider (2025b). The guy who coined 'vibe coding' predicts it will 'terraform software and alter job descriptions'. https://www.businessinsider.com/andrej-karpathy-coined-vibecoding-ai-prediction-2025-12

Cambridge (2025). Navigating China's regulatory approach to generative artificial intelligence and large language models. https://www.cambridge.org/core/journals/cambridge-forum-on-ai-law-and-governance/article/navigating-chinas-regulatory-approach-to-generative-artificial-intelligence-and-large-language-models/969B2055997BF42DE693B7A1A1B4E8BA

Centre for European Policy (2026). Competition in Generative AI: Updated Assessment. cepInput No. 1/2026. https://www.cep.eu/fileadmin/user_upload/cep.eu/Studien/cepInput_Competition_in_Generative_AI/cepInput_Competition_in_GenAI_Updated_Assessment.pdf

Chatgptiseatingtheworld (2026). Updated Master chart of copyright, DMCA and other claims in suits v. AI (Dec. 5, 2025). https://chatgptiseatingtheworld.com/2025/12/03/updated-master-chart-of-copyright-dmca-and-other-claims-in-suits-v-ai-dec-3-2025/

Chen (2025). Need Help? Designing Proactive AI Assistants for Programming. CHI 2025. ACM. https://doi.org/10.1145/3706598.3714002

Cheong (2025). E2E Process Automation Leveraging Generative AI and IDP-Based Automation Agent: A Case Study on Corporate Expense Processing. https://arxiv.org/abs/2505.20733

Christensen (1997). *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business School Press.

CKGSB (2025). Cheung Kong Graduate School of Business. Banking on data: How MYbank is revolutionizing supply chain finance. CKGSB Knowledge. https://english.ckgsb.edu.cn/knowledge/article/unleashing-innovation-in-china-series-banking-on-data-how-mybank-is-revolutionizing-supply-chain-finance/

Corrêa (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10), 100857. https://doi.org/10.1016/j.patter.2023.100857

Covington (2025). New Artificial Intelligence Legislation in Mexico. Global Policy Watch. https://www.globalpolicywatch.com/2025/03/new-artificial-intelligence-legislation-in-mexico/

CrowdStrike (2025). CrowdStrike Advances Next-Gen SIEM with Threat Hunting Across Data Sources, AI-Driven UEBA. https://www.crowdstrike.com/en-us/blog/crowdstrike-advances-next-gen-siem-capabilities/

Cyberhaven (2025). AI Adoption and Risk Report Q2 2025. https://info.cyberhaven.com/hubfs/Content %20PDF/Cyberhaven %20Labs %20- %202025 %20AI %20Adoption %20& %20Risk %20Report.pdf

Darktrace (2025). New Report Finds that 78 % of Chief Information Security Officers Globally are Seeing a Significant Impact from AI-Powered Cyber Threats – up 5 % from last year. https://www.darktrace.com/news/new-report-finds-that-78-of-chief-information-security-officers-globally-are-seeing-a-significant-impact-from-ai-powered-cyber-threats

DBS Bank (2024). DBS AI-Powered Digital Transformation. https://www.dbs.com/artificial-intelligence-machine-learning/artificial-intelligence/dbs-ai-powered-digital-transformation.html

Dealroom (2025). AI startups: Revenue per employee benchmarks. https://x.com/dealroomco/status/1914264599505018989

Deutsche Bank (2025). Claudio de Sanctis, Head of Private Bank, Deutsche Bank AG Private Bank. Investor Deep Dive 2025. https://investor-relations.db.com/files/documents/other-presentations-and-events/2025/IDD-2025-Script-Private-Bank-Claudio-de-Sanctis.pdf

DHL (2024). DHL Supply Chain Passes Unprecedented 500 Million Picks Milestone Using Locus Robotics Autonomous Mobile Robots. https://www.dhl.com/es-en/home/press/press-archive/2024/dhl-supply-chain-passes-unprecedented-500-million-picks-milestone-using-locus-robotics-autonomous-mobile-robots.html

EBA (2021). EBA Discussion Paper on Machine Learning for IRB Models. https://www.eba.europa.eu/sites/default/files/document_library/Publications/Discussions/2022/Discussion %20on %20machine %20learning %20for %20IRB %20models/1023883/Discussion %20paper %20on %20machine %20learning %20for %20IRB %20models.pdf

ECB (2025). ECB Guide to Internal Models. https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guide202507.en.pdf

EDPB (2025). AI Privacy Risks & Mitigations Large Language Models (LLMs). https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-privacy-risks-mitigations-large_en

Epoch (2025a). AI Benchmarking. https://epoch.ai/benchmarks

Epoch (2025b). How much power will frontier AI training demand in 2030? https://epoch.ai/blog/power-demands-of-frontier-ai-training

ESOMAR (2024). Global Market Research 2024. https://shop.esomar.org/knowledge-center/library?publication=3019

Eurostat (2025). 32.7 % of EU people used generative AI tools in 2025. https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20251216-3

Financial News London (2025). Deutsche Bank to roll out 'banking butlers' for affluent clients. https://www.fnlondon.com/articles/deutsche-bank-to-roll-out-banking-butlers-for-ultra-wealthy-clients-77e0349a

Figure (2025). F.02 Contributed to the Production of 30,000 Cars at BMW. https://www.figure.ai/news/production-at-bmw

FirstPageSage (2025). ChatGPT Usage Statistics: December 2025. https://firstpagesage.com/seo-blog/chatgpt-usage-statistics/

Fortune (2025a). Deloitte allegedly cited AI-generated research in a million-dollar report for a Canadian provincial government. https://fortune.com/2025/11/25/deloitte-caught-fabricated-ai-generated-research-million-dollar-report-canada-government/

Fortune (2025b). Elon Musk reveals massive plans for Tesla and Optimus—'Things are really going to go ballistic next year'. https://fortune.com/2025/01/30/elon-musk-reveals-massive-plans-tesla-optimus-self-driving-cars-humanoid-robots/

Fortune (2025c). AI enabled Klarna to halve its workforce—now, the CEO is warning other tech leaders to be honest about the risks. https://fortune.com/2025/10/10/klarna-ceo-sebastian-siemiatkowski-halved-workforce-says-tech-ceos-sugarcoating-ai-impact-on-jobs-mass-unemployment-warning/

Gartner (2025a). Hype Cycle for Artificial Intelligence. https://www.gartner.com/en/newsroom/press-releases/2025-08-05-gartner-hype-cycle-identifies-top-ai-innovations-in-2025

Gartner (2025b). Gartner Predicts Over 40 % of Agentic AI Projects Will Be Canceled by End of 2027. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

Google (2023). Practitioners Guide to MLOps: A framework for continuous delivery and automation of machine learning. https://cloud.google.com/resources/mlops-whitepaper

Google DeepMind (2025). AlphaFold: Science and impact. https://deepmind.google/science/alphafold/

Google Quantum AI (2024). Quantum error correction below the surface code threshold. *Nature*, 638, 920–926. https://doi.org/10.1038/s41586-024-08449-y

Grieves-Vickers (2017). Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. En Kahlen, F.J. et al. (eds.), *Transdisciplinary Perspectives on Complex Systems*. Springer. https://doi.org/10.1007/978-3-319-38756-7_4

Gu (2025). Large Language Models for Constructing and Optimizing Machine Learning Workflows: A Survey. https://arxiv.org/html/2411.10478v1

Heydari (2025). Tiny Machine Learning and On-Device Inference: A Survey of Applications, Challenges, and Future Directions. Sensors, 25(10), 3191. https://doi.org/10.3390/s25103191

HelpNetSecurity (2025). 67 % of daily security alerts overwhelm SOC analysts. https://www.helpnetsecurity.com/2023/07/20/soc-analysts-tools-effectiveness/

Hernández-Torres (2025). Challenges and Opportunities of Ambient Intelligence (AmI) in the 21st Century: A Historical Review. Evolutionary Intelligence, 18, 80. https://doi.org/10.1007/s12065-025-01010-z

Huang (2021). Power of data in quantum machine learning. *Nature Communications*, 12, 2631. https://doi.org/10.1038/s41467-021-22539-9

Hyundai (2025). Hyundai Motor Group to Unveil AI Robotics Strategy at CES 2026. https://www.hyundai.com/worldwide/en/newsroom/detail/0000001093

IBM (2025). Cost of a Data Breach Report 2025. https://www.ibm.com/reports/data-breach

IBM (2025b). Chief AI Officers cut through complexity to create new paths to value. https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/chief-ai-officer

Index Ventures (2026). Life, the Universe, and Simile: Leading Simile's $100M Series A. https://www.indexventures.com/perspectives/life-the-universe-and-simile-leading-similes-series-a/

iDanae (3Q20). Cátedra iDanae (UPM–Management Solutions). MLOps, a key element in the digital ecosystem. 2020. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2020/10/Idanae-3Q20.pdf

iDanae (2Q23). Cátedra iDanae (UPM–Management Solutions). Large Language Models: a new era for Artificial Intelligence. 2023. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2023/07/Idanae-2Q23.pdf**iDanae (1Q24).** Cátedra iDanae (UPM–Management Solutions). Towards a sustainable Artificial Intelligence. 2024. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2024/04/Idanae-1Q24-VDef.pdf

iDanae (1Q25). Cátedra iDanae (UPM–Management Solutions). The challenge of biases in the construction of Artificial Intelligence systems. 2025. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2025/04/Idanae-1Q25.pdf

iDanae (2Q25). Cátedra iDanae (UPM–Management Solutions). GenAI: an approach to multi-agents systems. 2025. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2025/07/Idanae-2Q25.pdf

IEA (2025a). Electricity 2025: Analysis and forecast to 2027. International Energy Agency. https://www.iea.org/reports/electricity-2025

IEA (2025b). World Energy Outlook Special Report on Energy and AI. International Energy Agency. https://www.iea.org/reports/energy-and-ai

IMD (2023). In the Field with Ping An. *IMD Business School*. https://www.imd.org/research-knowledge/digital/articles/in-the-field-with-ping-an/

IMF (2024). Gen-AI: Artificial Intelligence and the Future of Work. https://www.imf.org/-/media/files/publications/sdn/2024/english/sdnea2024001.pdf

ILO (2025a). International Labour Organization. Generative AI and Jobs: A Global Analysis of Potential Effects on Job Quantity and Quality. https://www.ilo.org/sites/default/files/2025-05/WP140_web.pdf

ILO (2025b). International Labour Organization. Governing AI in the World of Work: A review of global ethics guidelines. https://www.ilo.org/resource/article/governing-ai-world-work-review-global-ethics-guidelines

ILO (2025c). International Labour Organization. Global Case Studies of Social Dialogue on AI and Algorithmic Management. https://www.ilo.org/sites/default/files/2025-07/wp144_web.pdf

Inter-Parliamentary Union (2025). Parliamentary actions on AI policy. https://www.ipu.org/impact/democracy-and-strong-parliaments/artificial-intelligence/parliamentary-actions-ai-policy

Ironscales (2025). Ironscales Fall 2025 Threat Report. https://ironscales.com/hubfs/Landing %20Page %20Assets/Fall %202025 %20Threat %20Report/2025 %20Fall %20Threat %20Report_Beyond %20Detection_Reality %20of %20Deepfake %20Attacks %202.pdf

ISO/IEC (2023). ISO/IEC 42001:2023 - Artificial Intelligence Management Systems. https://www.iso.org/standard/42001

Jones (2025). Large Language Models Pass the Turing Test. https://arxiv.org/abs/2503.23674

Jumpcloud (2025). How Effective Is AI for Cybersecurity Teams? 2025 Statistics. https://jumpcloud.com/blog/how-effective-is-ai-for-cybersecurity-teams

KnowBe4 (2025). Phishing Threat Trends Report. https://www.knowbe4.com/hubfs/Phishing-Threat-Trends-2025_Report.pdf

Krijger, J. (2023). Operationalising ethics for AI in the financial industry: Insights from the Volksbank case study. *Journal of Digital Banking*, 8(3), 220–241. https://doi.org/10.69554/YQZC2796

Kumar (2020). Adversarial Machine Learning - A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2023. https://csrc.nist.gov/pubs/ai/100/2/e2023/final

Kusumegi et al. (2025). Scientific production in the era of large language models. Science. https://www.science.org/doi/10.1126/science.adw3000

Li (2025). Pitfalls and prospects of quantum machine learning. *Nature Computational Science*, 5, 1095–1097. https://doi.org/10.1038/s43588-025-00914-6

Liu (2024). Towards provably efficient quantum algorithms for large-scale machine-learning models. *Nature Communications*, 15, 434. https://doi.org/10.1038/s41467-023-43957-x

Lukac (2025). Ambient AI Scribes in Clinical Practice: A Randomized Trial. NEJM AI. https://doi.org/10.1056/AIoa2501000

Management Solutions (2023). Explainable Artificial Intelligence (XAI): Challenges of model interpretability. 2023. https://www.managementsolutions.com/en/microsites/whitepapers/explainable-artificial-intelligence

Management Solutions (3Q23). Follow-up report on Machine Learning for IRB models. Technical note on regulations, 3Q23, 2023.

https://www.managementsolutions.com/en/publications-and-events/regulatory-notes/technical-notes-on-regulations/follow-report-machine-learning-irb-models

Management Solutions (4Q24). Artificial Intelligence Act. Technical note on regulations, 4Q24, 2024. https://www.managementsolutions.com/en/publications-and-events/regulatory-notes/technical-notes-on-regulations/proposal-regulation-european-approach-artificial-intelligence

Management Solutions (4Q24a). Artificial Intelligence: regulatory landscape. Technical note on regulations, 4Q24, 2024. https://www.managementsolutions.com/en/publications-and-events/regulatory-notes/technical-notes-on-regulations/artificial-intelligence-regulatory-landscape

Marwala (2025). Why the Need for Governing Ambient Intelligence Has Never Been More Urgent. United Nations University. https://unu.edu/article/why-need-governing-ambient-intelligence-has-never-been-more-urgent

Mascelli (2025). Harvest Now Decrypt Later: Examining Post-Quantum Cryptography and the Data Privacy Risks for Distributed Ledger Networks. *Finance and Economics Discussion Series* 2025-093. Board of Governors of the Federal Reserve System. https://doi.org/10.17016/FEDS.2025.093

Microsoft (2026). AI-powered SIEM, built for modern security. https://marketingassets.microsoft.com/gdc/gdccKuCLQ/original

MIT Technology Review (2024). What's next for AlphaFold: A conversation with a Google DeepMind Nobel laureate. https://www.technologyreview.com/2025/11/24/1128322/whats-next-for-alphafold-a-conversation-with-a-google-deepmind-nobel-laureate/

MITRE (2025). ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems). https://atlas.mitre.org/

Moroni (2025). Insurmountable Limitations of City-Scale Digital Twins? On Urban Knowledge and Planning. Computational Urban Science. Springer Nature. https://doi.org/10.1007/s43762-025-00174-0

Nadella (2025). LinkedIn Post. https://www.linkedin.com/posts/satyanadella_just-wrapped-our-earnings-call-and-wanted-activity-7389433821181562880-GnMz/

Nissenbaum (1996). Accountability in a Computerized Society. Science and Engineering Ethics, 2, 25–42. https://link.springer.com/article/10.1007/BF02639315

NIST (2023). AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework

NIST (2024a). Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile. https://www.nist.gov/publications/secure-software-development-practices-generative-ai-and-dual-use-foundation-models-ssdf

NIST (2024b). Post-Quantum Cryptography Standards: FIPS 203, FIPS 204, FIPS 205. National Institute of Standards and Technology. https://csrc.nist.gov/projects/post-quantum-cryptography

Notateslaapp (2025). Tesla Eyes $20K Price Target For Optimus, Extremely Fast Production Ramp. https://www.notateslaapp.com/news/3314/tesla-eyes-20k-price-target-for-optimus-extremely-fast-production-ramp

OECD (2024). A Sectoral Taxonomy of AI Intensity. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/12/a-sectoral-taxonomy-of-ai-intensity_c2baae71/1f6377b5-en.pdf

OECD (2025a). Emerging Divides in the Transition to Artificial Intelligence. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/emerging-divides-in-the-transition-to-artificial-intelligence_eeb5e120/7376c776-en.pdf

OECD (2025b). The effects of generative AI on productivity, innovation and entrepreneurship. OECD Artificial Intelligence Papers, no. 39. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/the-effects-of-generative-ai-on-productivity-innovation-and-entrepreneurship_da1d085d/b21df222-en.pdf

OECD (2026). AI Use by Individuals Surges Across the OECD as Adoption by Firms Continues to Expand. https://www.oecd.org/en/about/news/announcements/2026/01/ai-use-by-individuals-surges-across-the-oecd-as-adoption-by-firms-continues-to-expand.html

Ouyang (2025). FELA: A Multi-Agent Evolutionary System for Feature Engineering of Industrial Event Log Data. https://arxiv.org/html/2510.25223

OWASP (2025). OWASP Top 10 for Large Language Model Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/

Oxford (2025). AI Regulation: The Politics of Fragmentation and Regulatory Capture. https://blogs.law.ox.ac.uk/oblb/blog-post/2025/06/ai-regulation-politics-fragmentation-and-regulatory-capture

Parfit (1984). Reasons and Persons. Oxford University Press.

Park (2023). Generative Agents: Interactive Simulacra of Human Behavior. UIST '23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. ACM. https://doi.org/10.1145/3586183.3606763

Park (2024). Generative Agent Simulations of 1,000 People. https://arxiv.org/abs/2411.10109

Patel (2025). U.S. AI Law & Policy Explained. Enterprise AI Governance. https://oliverpatel.substack.com/p/us-ai-law-and-policy-explained

Pew (2025). Pew Research Institute. How the U.S. Public and AI Experts View Artificial Intelligence. https://www.pewresearch.org/

Phishcare (2025). Top 10 Deepfake Phishing Scams. https://phishcare.com/top-10-deepfake-phishing-scams/

Pixiebrix (2025). Top Chief AI Officers of 2025. https://www.pixiebrix.com/reports/top-ai-officers-of-2025

PR Newswire (2023). SlashNext's 2023 State of Phishing Report Reveals a 1,265 % Increase in Phishing Emails Since the Launch of ChatGPT in November 2022, Signaling a New Era of Cybercrime Fueled by Generative AI. https://www.prnewswire.com/news-releases/slashnexts-2023-state-of-phishing-report-reveals-a-1-265-increase-in-phishing-emails-since-the-launch-of-chatgpt-in-november-2022--signaling-a-new-era-of-cybercrime-fueled-by-generative-ai-301971557.html

Proofpoint (2025). AI Threat Detection. https://www.proofpoint.com/us/threat-reference/ai-threat-detection

Pu (2025). Assistance or Disruption? Exploring and Evaluating the Design and Trade-offs of Proactive AI Programming Support. CHI 2025. ACM. https://doi.org/10.1145/3706598.3713357

Quartz (2025). AI powers smaller startups toward a new era of unicorns. https://qz.com/unicorn-entrepreneur-founder-solo-ai-startup-automation-workforce

Rawls (1971). A Theory of Justice. Harvard University Press.

Ryt Bank (2025). The World's First AI-Powered Bank. https://www.rytbank.my/

Sacra (2026). Cursor revenue, valuation & funding. https://sacra.com/c/cursor/

Shan (2024). Transitioning from MLOps to LLMOps: Navigating the Unique Challenges of Large Language Models. https://doi.org/10.3390/info16020087

Shankar (2021). Towards Observability for Machine Learning Pipelines. DOI:10.48550/arXiv.2108.13557. (PDF) Towards Observability for Machine Learning Pipelines

SoSafe (2025). Global businesses face escalating AI risk, as 87 % hit by AI cyberattacks. https://sosafe-awareness.com/company/press/global-businesses-face-escalating-ai-risk-as-87-hit-by-ai-cyberattacks/

SQ Magazine (2025). AI Cyber Attacks Statistics 2025: How Attacks, Deepfakes & Ransomware Have Escalated. https://sqmagazine.co.uk/ai-cyber-attacks-statistics/

Stanford (2023a). Stanford Encyclopedia of Philosophy. Ethics of Artificial Intelligence and Robotics. https://plato.stanford.edu/entries/ethics-ai/

Stanford (2023b). Stanford Encyclopedia of Philosophy. Philosophy of Artificial Intelligence. https://plato.stanford.edu/entries/artificial-intelligence/

Stanford (2025). The 2025 AI Index Report. https://hai.stanford.edu/ai-index/2025-ai-index-report

Stone (2025). Navigating MLOps: Insights into Maturity, Lifecycle, Tools, and Careers. https://arxiv.org/html/2503.15577v1

Tesla Car World (2025). Elon Musk Unveils Tesla Bot Gen 3 Real Homemaker Updates. https://www.youtube.com/watch?v=HR1HrrneNHs&t=1s

Teslarati (2025). Tesla Optimus' pilot line will already have an incredible annual output. https://www.teslarati.com/tesla-optimus-pilot-line-will-already-have-an-incredible-annual-output/

The Network Installers (2025). AI Cyber Threat Statistics. https://thenetworkinstallers.com/es/blog/ai-cyber-threat-statistics/

Thompson (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. American Political Science Review, 74(4), 905–916. https://www.cambridge.org/core/journals/american-political-science-review/article/abs/moral-responsibility-of-public-officials-the-problem-of-many-hands/39DD3FAB7BF7DC7A242407143674F22B

Toyota Research Institute (2025). AI-Powered Robot by Boston Dynamics and Toyota Research Institute Takes a Key Step Towards General-Purpose Humanoids. https://www.tri.global/news/ai-powered-robot-boston-dynamics-and-toyota-research-institute-takes-key-step-towards-general

UK AI Safety Institute (2026). International AI Safety Report 2026. UK Government. https://www.gov.uk/government/publications/international-ai-safety-report-2026

UK DSIT (2023). UK Department for Science, Innovation and Technology. Public Attitudes to Data and AI: Tracker Survey (Report). https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey

UK Government (2023). A pro-innovation approach to AI regulation. Policy paper. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

UNDP (2025). Human Development Report 2025. United Nations Development Programme. https://hdr.undp.org/content/human-development-report-2025

UNESCO (2023). Guidance for Generative AI in Education and Research. https://unesdoc.unesco.org/ark:/48223/pf0000386693

United Nations (2025). The Sustainable Development Goals Report 2025. United Nations, Department of Economic and Social Affairs. https://unstats.un.org/sdgs/report/2025/

WatchGuard (2025). Evasive Malware Surges 40 % in WatchGuard's Latest Internet Security Report. https://www.watchguard.com/es/wgrd-news/blog/evasive-malware-surges-40-watchguards-latest-internet-security-report

WIPO (2025). WIPO Conversation on Intellectual Property and Frontier Technologies. https://www.wipo.int/en/web/frontier-technologies/frontier_conversation

World Bank (2024). Global Trends in AI Governance. https://documents1.worldbank.org/curated/en/099120224205026271/pdf/P1786161ad76ca0ae1ba3b1558ca4ff88ba.pdf

World Bank (2025). World Development Report 2025: Standards for Development. World Bank. https://www.worldbank.org/en/publication/wdr2025

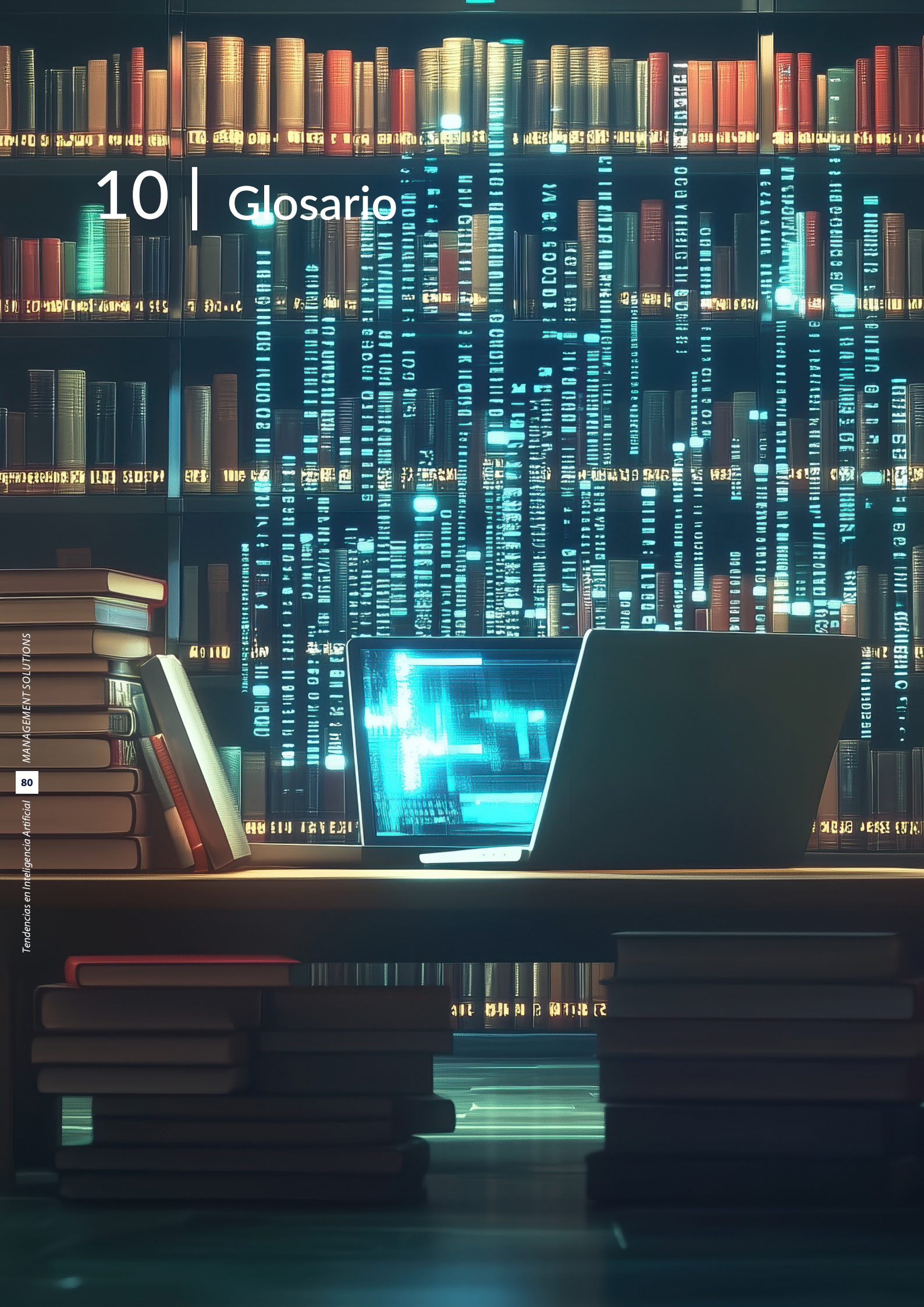
World Economic Forum (2025a). AI in Action: Beyond Experimentation to Transform Industry. https://reports.weforum.org/docs/WEF_AI_in_Action_Beyond_Experimentation_to_Transform_Industry_2025.pdf

World Economic Forum (2025b). Transforming Consumer Industries in the Age of AI. https://reports.weforum.org/docs/WEF_Transforming_Consumer_Industries_in_the_Age_of_AI_2025.pdf

World Health Organization (2024). Ethics and Governance of Artificial Intelligence for Health. https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content

Yuan et al. (2025). The Impact of AI Adoption in the Workplace on Employees: A Systematic Review. https://www.researchgate.net/publication/396219396_The_Impact_of_AI_Adoption_in_the_Workplace_on_Employees_A_Systematic_Review

10 | Glosario



AGI (Artificial General Intelligence): Sistema de IA capaz de realizar cualquier tarea cognitiva que pueda realizar un ser humano, con generalización transferible entre dominios. Horizonte estratégico debatido cuya definición precisa carece de consenso científico.

AI Act: Reglamento (UE) 2024/1689, primer marco legal integral sobre IA. Clasifica los sistemas por nivel de riesgo (inaceptable, alto, limitado, mínimo) e impone obligaciones estructurales a los de alto riesgo. Sanciones de hasta 35 M€ o el 7 % de la facturación global.

AI Board: Foro de coordinación e interpretación común entre la Comisión Europea y las autoridades nacionales de supervisión de IA, creado por el AI Act.

AI Office: Órgano técnico central de la Comisión Europea responsable de la supervisión de modelos de IA de propósito general (GPAI) en el marco del AI Act.

AI-enhanced: Modelo organizativo en el que la IA se usa para optimizar procesos existentes sin rediseñarlos desde las capacidades de la IA. Estadio predominante en la mayoría de organizaciones actuales.

AI-first: Modelo organizativo en el que los procesos y la estructura se diseñan partiendo de las capacidades de la IA, asignando al juicio humano únicamente las tareas donde su ventaja comparativa es inequívoca.

AI-only: Modelo organizativo hipotético en el que las operaciones centrales prescinden funcionalmente del trabajo humano.

AIMS (AI Management System): Sistema de gestión de IA conforme a ISO/IEC 42001, equivalente a ISO 27001 para ciberseguridad pero específico para IA: cubre políticas, evaluaciones de impacto, control de proveedores y supervisión continua.

Alucinación: Comportamiento intrínseco de los modelos generativos por el que producen información falsa o inexacta presentada con aparente confianza. No es un bug ocasional, sino una consecuencia estructural del diseño actual de los LLMs.

Ambient AI (IA ambiental): IA que opera sin ser invocada explícitamente: observa el contexto de forma continua, infiere necesidades y actúa de forma proactiva. La interfaz desaparece; el entorno mismo se convierte en el punto de interacción.

Ambient scribe: Sistema de Ambient AI desplegado en entornos clínicos que escucha la conversación médico-paciente y genera automáticamente la documentación del encuentro sin instrucción explícita del usuario.

BEC (Business Email Compromise): Tipo de ciberataque en el que se suplanta la identidad de un directivo o proveedor para desviar fondos o extraer información. La IA generativa lo potencia mediante deepfakes de audio y vídeo de alta credibilidad.

Brecha de absorción: Distancia creciente entre la curva de capacidad tecnológica de la IA (exponencial, autoacelerada) y la curva de absorción organizativa (rediseño de procesos, reconversión de roles, gobernanza).

CAIO / CDAIO (Chief AI Officer / Chief Data and AI Officer): Función ejecutiva responsable del liderazgo estratégico de la IA en una organización.

Citizen data scientist: Profesional no técnico capaz de realizar análisis de datos básicos con herramientas visuales. La IA generativa supera este concepto al entregar capacidades analíticas avanzadas directamente a usuarios finales sin formación técnica.

Computación cuántica: Paradigma computacional que explota propiedades cuánticas (superposición, entrelazamiento) para resolver ciertos problemas intratables para sistemas clásicos.

Criptografía resistente a ataques cuánticos (PQC): Conjunto de algoritmos criptográficos diseñados para resistir ataques de ordenadores cuánticos. El NIST publicó los primeros estándares en 2024 (FIPS 203, 204, 205).

Dark LLM: Modelos de lenguaje modificados específicamente para cibercrimen (WormGPT, FraudGPT, GhostGPT). Generan malware, exploits y campañas de ingeniería social sin restricciones éticas. Comercializados en dark web con soporte técnico.

Data poisoning: Ataque adversario que consiste en inyectar datos maliciosos en el conjunto de entrenamiento de un modelo para degradar su comportamiento o introducir sesgos controlados por el atacante.

Deepfake: Contenido audiovisual sintético generado por IA que suplanta la apariencia o voz de una persona real. Usado en ataques BEC, fraude y desinformación con tasa de éxito significativamente superior al phishing tradicional.

Differential privacy: Técnica que añade ruido estadístico controlado a los datos o resultados para impedir la identificación de individuos, preservando utilidad estadística agregada.

DPIA (Data Protection Impact Assessment): Evaluación de impacto en protección de datos exigida por GDPR (Art. 35). El EDPB la considera obligatoria en la mayoría de despliegues de LLMs dado su procesamiento sistémico de datos personales.

Edge AI / TinyML: Capacidad de ejecutar modelos de IA directamente en dispositivos como smartphones, wearables y sensores sin depender de conectividad con la nube. Habilitador clave de la Ambient AI.

Efecto Bruselas: Fenómeno por el que la regulación de la UE (AI Act, GDPR) obliga a empresas globales a adaptar sus productos a los estándares europeos por el volumen del mercado, exportando de facto esa regulación al resto del mundo.

Ethics washing: Fenómeno por el que una organización publica principios éticos de IA sin traducirlos en controles operativos, responsabilidades concretas o mecanismos de auditoría.

Evasión adversaria: Ataque que introduce perturbaciones imperceptibles en los inputs de un modelo para provocar clasificaciones o respuestas incorrectas durante la inferencia, sin modificar el modelo en sí.

Explicabilidad: Capacidad de describir el funcionamiento interno de un modelo de IA de forma comprensible para diferentes audiencias (regulador, cliente, empleado). Requisito técnico en modelos regulados; implementable en ML con técnicas como SHAP, LIME o análisis de sensibilidad.

Feature engineering: Proceso de construir variables predictivas a partir de datos crudos para alimentar modelos de ML. Una de las fases más intensivas en conocimiento experto del ciclo de vida del ML clásico; acelerada significativamente por IA generativa.

Federated learning: Paradigma de entrenamiento distribuido en el que los datos permanecen en los dispositivos locales y solo se comparten actualizaciones del modelo. Mitiga riesgos de privacidad en LLMs a costa de mayor complejidad operativa.

GDPR (General Data Protection Regulation): Reglamento (UE) 2016/679 de protección de datos. Sus principios de minimización, derecho al olvido y transparencia presentan tensiones estructurales con la arquitectura y el ciclo de vida de los LLMs.

Gemelo digital (Digital Twin): Simulación dinámica de un sistema físico o humano que se actualiza en tiempo real con datos del sistema real. Funciona con alta fiabilidad en sistemas físicos deterministas; los LLMs han abierto su extensión a comportamiento humano y sistemas sociales complejos.

GPAI (General Purpose AI): Modelo de IA de propósito general (e.g., GPT, Claude, Gemini) capaz de realizar una amplia variedad de tareas.

Gradient boosting: Técnica de ML que combina múltiples árboles de decisión de forma secuencial para mejorar la predicción.

Harvest now, decrypt later: Estrategia por la que actores estatales capturan hoy comunicaciones cifradas con la intención de descifrarlas cuando la computación cuántica madure. Amenaza presente para datos con larga vida útil de confidencialidad.

Hub & spokes: Modelo organizativo de IA con un Centro de Excelencia central que establece capacidades transversales, mientras equipos descentralizados en líneas de negocio desarrollan soluciones específicas con reporte funcional cruzado al hub.

IA agéntica: Sistemas de IA que planifican, ejecutan tareas complejas y operan de forma autónoma sobre infraestructuras corporativas reales, más allá de responder a prompts. Capacidades incrementales: estado persistente, planificación dinámica, ejecución sobre sistemas reales y orquestación multiagente.

IA generativa: Familia de modelos capaces de generar contenido original (texto, imágenes, audio, vídeo, código) ante instrucciones en lenguaje natural. Basada en arquitecturas transformer de gran escala.

IRB (Internal Ratings-Based): Enfoque regulatorio que permite a las entidades bancarias usar modelos internos para calcular el capital regulatorio por riesgo de crédito.

Jagged intelligence (inteligencia dentada): Concepto de Andrej Karpathy para describir el perfil de capacidades de los LLMs actuales: brillantes en tareas complejas (olimpiadas matemáticas) y frágiles en tareas aparentemente simples (comparar números decimales).

Jailbreak: Técnica para eludir los controles de seguridad de un modelo de IA mediante instrucciones diseñadas para hacer que el sistema ignore sus restricciones de comportamiento.

LIME (Local Interpretable Model-agnostic Explanations): Técnica de explicabilidad que genera aproximaciones locales de un modelo complejo para explicar predicciones individuales.

LLM (Large Language Model): Modelo de lenguaje de gran escala basado en arquitectura transformer, entrenado sobre vastos corpus textuales. Fundamento técnico de parte de la IA generativa actual. Ejemplos: GPT, Claude, Gemini, Llama.

LLMOps (Large Language Model Operations): Extensión de MLOps específica para gestionar las propiedades de los LLMs en producción: comportamiento no determinista, prompts como superficie de riesgo, alucinaciones, costes por token y trazabilidad de interacciones.

Lock-in de proveedor: Dependencia estructural de un único proveedor de modelos o infraestructura que dificulta la migración (reescritura de integraciones, recertificación de compliance, costes prohibitivos). Riesgo estratégico en la adopción de IA.

LoRA (Low-Rank Adaptation): Técnica eficiente de fine-tuning de LLMs que adapta el modelo base a un dominio específico modificando solo un subconjunto de parámetros, reduciendo drásticamente los recursos computacionales necesarios.

Machine learning (ML): Rama de la IA que desarrolla algoritmos capaces de aprender patrones a partir de datos históricos sin ser programados explícitamente para cada tarea. A diferencia de la IA generativa, los modelos de ML clásico clasifican, predicen y optimizan; no generan contenido.

Machine unlearning: Conjunto de técnicas experimentales que buscan eliminar selectivamente el conocimiento derivado de datos específicos de un modelo ya entrenado, en respuesta al derecho al olvido del GDPR.

Malware polimórfico: Código malicioso que muta su firma para evadir detección. La IA generativa lo potencia: variantes avanzadas reescriben su código cada 15 segundos manteniendo funcionalidad idéntica, y vencen a sistemas basados en firmas estáticas.

Many hands problem: Problema de responsabilidad difusa en sistemas complejos donde el daño se produce sin intención deliberada y la cadena causal se fragmenta entre múltiples actores (diseñadores, entrenadores, desplegadores, usuarios).

MCP (Model Context Protocol): Estándar abierto que normaliza cómo los modelos de IA interactúan con aplicaciones, fuentes de datos y herramientas externas. Elimina la deuda técnica de las integraciones propietarias y convierte cada herramienta en un activo reutilizable por cualquier agente.

MLOps (Machine Learning Operations): Conjunto de procesos estandarizados y capacidades tecnológicas para construir, desplegar y operacionalizar modelos de ML de forma fiable a lo largo de todo su ciclo de vida. Cubre preparación de datos, experimentación, validación, despliegue y monitorización.

Model drift: Degradación silenciosa del rendimiento de un modelo en producción causada por cambios en la distribución de los datos de entrada respecto a los de entrenamiento.

Multimodalidad: Capacidad de un modelo de IA de procesar y generar simultáneamente múltiples tipos de información (texto, imágenes, audio, vídeo, código) en una sola arquitectura conversacional.

NIST AI RMF: Marco de gestión de riesgos de IA del National Institute of Standards and Technology. Organizado en las funciones Govern-Map-Measure-Manage. Orientado a características de IA fiable (trustworthy AI).

Orquestación multiagente: Arquitectura en la que múltiples agentes de IA especializados colaboran bajo un coordinador central para alcanzar objetivos complejos, gestionando dependencias, prioridades y traspaso de información entre agentes.

Phishing hiperpersonalizado: Ataques de phishing generados por IA que analizan perfiles públicos y estilo de escritura corporativa para crear mensajes altamente personalizados.

Prompt: Instrucción o entrada en lenguaje natural que un usuario proporciona a un sistema de IA generativa para obtener una respuesta.

Prompt injection: Ataque en el que instrucciones maliciosas embebidas en los inputs (documentos, URLs, datos externos) manipulan el comportamiento del modelo para ignorar restricciones o ejecutar acciones no autorizadas.

RAG (Retrieval-Augmented Generation): Arquitectura que complementa un LLM con un sistema de recuperación de información externa en tiempo real. Reduce alucinaciones al anclar las respuestas en documentos verificables, y separa el conocimiento actualizable de la memoria estática del modelo.

ReAct (Reason + Act): Bucle de control de agentes de IA que combina razonamiento (análisis de la situación) y acción (uso de herramientas o APIs) de forma iterativa. Fundamento del núcleo cognitivo de los sistemas agénticos.

Reskilling: Proceso de adquisición de nuevas competencias para desempeñar roles profesionales diferentes a los actuales, en respuesta a la transformación del trabajo por la IA. Palanca estratégica dado el desequilibrio estructural entre oferta y demanda de talento en IA.

Robótica humanoide: Robots con forma y capacidades de movimiento similares a las humanas, integrados con modelos de IA para percepción, razonamiento y aprendizaje.

Shadow AI: Uso no autorizado de herramientas de IA generativa en entornos corporativos, con frecuencia en plataformas públicas sin garantías de privacidad.

SHAP (SHapley Additive exPlanations): Técnica de explicabilidad basada en teoría de juegos que cuantifica la contribución de cada variable a una predicción individual de un modelo.

SIEM / XDR / SOAR: Plataformas de ciberseguridad: SIEM (Security Information and Event Management), XDR (Extended Detection and Response) y SOAR (Security Orchestration, Automation and Response).

Soberanía tecnológica: Capacidad de un estado u organización de controlar las capas críticas de la cadena de valor de la IA (hardware, infraestructura, modelos, talento) para mantener autonomía estratégica.

Tecnobloques: Esferas de influencia tecnológica parcialmente incompatibles (lideradas por EEUU, China y Europa) con lógicas propias de gobernanza, seguridad y valores.

Test de Turing: Criterio propuesto por Alan Turing (1950) para evaluar si una máquina exhibe comportamiento inteligente indistinguible del humano en conversación.

Transformer: Arquitectura de red neuronal escalable basada en mecanismos de atención, introducida por Google en 2017. Fundamento técnico de todos los LLMs modernos.

UEBA (User and Entity Behavior Analytics): Sistemas de ciberseguridad que establecen líneas base dinámicas de comportamiento normal y detectan anomalías.

Upskilling: Proceso de ampliación de competencias existentes para adaptarse a nuevos requisitos del mismo rol profesional, específicamente en el contexto de la adopción de IA.

Vibe coding: Paradigma de desarrollo de software por conversación iterativa en lenguaje natural con sistemas de IA que interpretan requisitos, generan aplicaciones completas, detectan errores y producen tests y documentación de forma automática. Término acuñado por Andrej Karpathy.

Nuestro objetivo es superar las expectativas de nuestros clientes convirtiéndonos en socios de confianza

Management Solutions es una firma internacional de servicios de consultoría centrada en el asesoramiento de negocio, finanzas, riesgos, organización y procesos, tanto en sus componentes funcionales como en la implantación de sus tecnologías relacionadas.

Con un equipo multidisciplinar (funcionales, matemáticos, técnicos, etc.) de más de 4.000 profesionales, Management Solutions desarrolla su actividad a través de 52 oficinas (22 en Europa, 24 en América, 3 en Asia, 1 en África y 2 en Oceanía).

Para dar cobertura a las necesidades de sus clientes, Management Solutions tiene estructuradas sus prácticas por industrias (Entidades Financieras, Energía, Telecomunicaciones y Otras industrias) y por líneas de actividad que agrupan una amplia gama de competencias: Estrategia, Gestión Comercial y Marketing, Gestión y Control de Riesgos, Información de Gestión y Financiera, Transformación: Organización, Procesos y Sistemas, y Nuevas Tecnologías y Metodologías.

Javier Calvo Martín
Socio y Chief AI Officer de Management Solutions
javier.calvo.martin@managementsolutions.com

Manuel Ángel Guzmán Caba
Socio de Management Solutions
manuel.guzman@managementsolutions.com

Segismundo Jiménez Laínez
Director de Management Solutions
segismundo.jimenez@managementsolutions.com

Louis Perron
Gerente de Management Solutions
louis.perron@managementsolutions.com

Management Solutions, servicios profesionales de consultoría

Management Solutions es una firma internacional de consultoría centrada en el asesoramiento de negocio, finanzas, riesgos, organización, tecnología y procesos.

Para más información visita **www.managementsolutions.com**

Síguenos en:

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña Málaga London Frankfurt Düsseldorf Wien Paris Bruxelles Amsterdam Copenhagen Oslo Stockholm Warszawa Wrocław Zürich Milano Roma Bologna Lisboa Beijing Abu Dhabi İstanbul Johannesburg Sydney Melbourne Toronto New York New Jersey Boston Pittsburgh Columbus Atlanta Birmingham Houston Phoenix Miami SJ de Puerto Rico San José Ciudad de México Monterrey Querétaro Medellín Bogotá Quito São Paulo Rio de Janeiro Lima Santiago de Chile Buenos Aires