Cuadernos de la Cátedra CaixaBank de Sostenibilidad e Impacto Social

Hacia una inteligencia artificial ética en la empresa: aplicaciones, riesgos y respuestas normativas para su fiabilidad

Bruno Martínez

Joan Fontrodona

N.º 62 - Noviembre del 2024

IESE Business School University of Navarra

Cátedra CaixaBank de Sostenibilidad e Impacto Social

Hacia una inteligencia artificial ética en la empresa: aplicaciones, riesgos y respuestas normativas para su fiabilidad

Bruno Martínez

Asistente de investigación

Joan Fontrodona

Profesor de Ética Empresarial y Análisis de Situaciones de Negocio y

titular de la Cátedra CaixaBank de Sostenibilidad e Impacto Social

Edición: Caja Alta Edición & Comunicación (www.cajaalta.es)

La Cátedra CaixaBank de Sostenibilidad e Impacto Social responde al compromiso de fomentar, promocionar y divulgar nuevos conocimientos sobre la Responsabilidad Social Empresarial (RSE). Todo esto, a través de la generación de ideas y conceptos innovadores en el ámbito de la responsabilidad social, dirigidos especialmente al sector empresarial.

Creada en 2005, la Cátedra desarrolla proyectos de investigación, participa en congresos y conferencias, y organiza mesas redondas y actividades de divulgación sobre la responsabilidad social de la empresa.

ÍNDICE

Introducción

La inteligencia artificial (IA), popularizada gracias a su uso mediante herramientas de IA generativa (GenIA) como ChatGPT-4, se sitúa en el centro del debate actual. Esta tecnología ha revolucionado múltiples sectores tanto por las posibilidades que ofrece como por los diversos desafíos éticos y preocupaciones que plantea. Es decir, la capacidad de estas herramientas generativas para producir contenido coherente y relevante en una gran variedad de contextos ha transformado el modo en que las empresas operan, los profesionales desarrollan sus labores y los individuos interactúan con la tecnología, pero, al mismo tiempo, también ha suscitado importantes debates acerca de la privacidad, la seguridad de los datos y su impacto en el empleo. Del mismo modo, la posibilidad que abre para propósitos maliciosos como la desinformación y la manipulación de la opinión pública ha llevado a expertos y legisladores a considerar la necesidad de establecer regulaciones claras y efectivas.

De hecho, en el año que en breve finalizaremos –2024–, se ha alcanzado un hito histórico con la aprobación del primer marco jurídico integral sobre este tecnología a nivel mundial: la Ley de IA de la Unión Europea (UE). Esta normativa forma parte de toda una serie de medidas políticas destinadas a apoyar y reforzar el desarrollo, la adopción, la inversión y la innovación en esta materia en toda la UE, con el fin de garantizar que esta sea fiable, es decir, que no atente contra la seguridad ni los derechos fundamentales de las personas y las empresas.

(…) el uso de la IA plantea toda una serie de desafíos éticos que también afectan a las empresas que hacen uso de ella (…)

Pese a su reciente popularización, la IA lleva mucho tiempo entre nosotros; no obstante, según estudios recientes, un 44% de los adultos estadounidenses encuestados en el 2022 aseguró no interactuar en su día a día con ella (Kennedy *et al.* 2023) y un 51% de la población mundial aseguraba, en el 2023, desconocer qué productos usaban esta tecnología (Ipsos 2023). Sin embargo, llevamos tiempo integrando herramientas de corrección de textos y de reconocimiento facial, filtros de *spam* para nuestros correos electrónicos o sistemas de recomendación, sin identificarlos propiamente como tal. De hecho, los conceptos y esfuerzos iniciales realizados en el ámbito de la IA dieron comienzo a mediados del siglo XX. Desde entonces, ha ido evolucionando desde simples algoritmos de aprendizaje automático o *machine learning (ML)* hasta las complejas redes neuronales y los modelos generativos actuales. A lo largo de este recorrido, ha sido integrada en áreas diversas como la medicina, la industria, la seguridad y el entretenimiento, mucho antes de la popularización de tecnologías actuales como la mencionada aplicación ChatGPT-4.

Parece innegable que, hoy en día, nos encontramos inmersos en una verdadera revolución de la IA que ha alterado nuestras vidas y a la que el sector empresarial no es ajeno. Según un reciente estudio de McKinsey (QuantumBlack, AI by McKinsey 2024), un 72% de las compañías encuestadas aseguró haber adoptado este año esta tecnología en, al menos, una función comercial, frente al 20% registrado en el 2017. Tal como veremos en este cuaderno, cada vez son más las organizaciones que integran sistemas y funcionalidades de IA en una gran diversidad de actividades: automatización de procesos, análisis de datos, atención al cliente, optimización logística, desarrollo de productos, marketing y publicidad, etc.

Así las cosas, no cabe duda de que, tal como se ha señalado, el uso de la IA plantea toda una serie de desafíos éticos que también afectan a las empresas que hacen uso de ella: la posibilidad de implementar un sistema que produzca discriminaciones, no sea justo, no respete la privacidad de las personas o no sea transparente en sus procedimientos puede afectar gravemente a las organizaciones y a su reputación en un panorama en el que la sociedad está demandando un nivel cada vez mayor de responsabilidad y ética. Por ello, para garantizar que el desarrollo de la IA sea beneficioso para todas las partes interesadas, resulta fundamental que las empresas también asuman su parte de responsabilidad en toda esta cuestión, llevando a cabo una selección de aquellos sistemas que garanticen una serie de principios de responsabilidad. Con ese fin, tal como veremos, tienen a su alcance la posibilidad de recurrir a una gran variedad de certificaciones y estándares que avalan los sistemas de IA a partir de toda una serie de consideraciones de carácter ético.

Este nuevo panorama en el que nos encontramos evoluciona a gran velocidad, por lo que resulta lógico que se tengan posiblemente más preguntas que respuestas. Con el fin de arrojar luz al respecto, el objetivo de este cuaderno es contextualizar la IA en el ámbito empresarial, resaltar algunas de las principales cuestiones relacionadas con ella que se plantean en la actualidad y apuntar algunas de las propuestas más relevantes que se han desarrollado o se están desarrollando para conseguirlo.

A nivel estructural, en primer lugar observaremos a grandes rasgos qué entendemos por IA, qué tipologías existen y cuáles están en fase de desarrollo, así como cuál es su estado actual. A continuación, veremos de forma concreta cómo está siendo integrada en el sector empresarial, así como los

principales ámbitos en los que las organizaciones la están incorporando. Más adelante, analizaremos los principales desafíos que plantea a nivel ético, tales como la presencia de sesgos y la discriminación que puede producir, el respeto a la privacidad de las personas o la suficiencia de transparencia para garantizar un uso responsable de ella, junto con diversas regulaciones y medidas se están implementando para abordar esos desafíos. Por último, mostraremos algunas opciones disponibles en la actualidad en el ámbito de las certificaciones y los estándares a los que las empresas pueden recurrir para asegurarse de que la implementación y el uso de sistemas de IA en su seno sean responsables y éticos.

La inteligencia artificial

Por *inteligencia artificial (IA)* se entiende aquel campo de la informática orientado a la creación de máquinas o programas capaces de realizar de forma autónoma tareas que, de otro modo, requerirían inteligencia y agencia humana (UNESCO 2021). Por tanto, hablamos de un sistema de IA para referirnos a un sistema informático con la capacidad de imitar funciones cognitivas humanas, como el aprendizaje y la solución de problemas (Microsoft Azure, s. f.). Se trata de una tecnología que, mediante el uso de las matemáticas y la lógica, simula el razonamiento de las personas para aprender a partir de información nueva y utilizar lo aprendido en la toma de decisiones futuras, del mismo modo que lo haría un ser humano.

Si bien todo lo relativo a la IA ha ganado un indudable gran peso en el debate actual, son múltiples los ejemplos de que hemos estado integrándola en nuestras vidas sin –seguramente– identificarlos como tal, y que, sin embargo, son ampliamente conocidos y utilizados en los diferentes sectores de la sociedad. Es el caso, por ejemplo, de los asistentes virtuales inteligentes –como Siri, de Apple; Google Assistant, Amazon Alexa, etc.–, los sistemas de reconocimiento facial o los de recomendación, los cuales, a partir de nuestro comportamiento como usuarios, nos recomiendan productos o contenido relevante. Sin embargo, ha sido a raíz de la aparición y popularización de la GenIA cuando se ha abierto un extenso debate acerca de sus límites éticos y sociales. Estos sistemas han demostrado tener una capacidad para generar textos, imágenes y música que, a menudo, resultan indistinguibles de los creados por personas, lo que ha puesto sobre la mesa importantes cuestiones acerca de la autoría, la responsabilidad y el impacto de estas tecnologías en diferentes ámbitos de la sociedad.

A la hora de identificar los distintos tipos de IA que existen, predominan dos clasificaciones, en función de la perspectiva desde la cual se observan. La primera clasificación tiene en cuenta sus capacidades, es decir, el grado en el que pueden imitar los procesos del pensamiento humano. Desde este enfoque se establecen tres tipos (Naveen 2019; IBM Data and AI Team 2023):

- **IA limitada:** la ANI *(artificial narrow intelligence)* es aquella IA que ha sido diseñada y entrenada para una tarea específica o para un rango reducido de tareas, bajo una serie de restricciones operativas, de modo que no puede desempeñarse fuera de ellas. En este sentido, no posee entendimiento, sino que sigue reglas preprogramadas y aprende patrones a partir de la información recogida. Es el caso, por ejemplo, de los sistemas de asistencia virtual personal como Alexa o Siri, de los sistemas de recomendación, de los *softwares* de reconocimiento de imágenes y de las herramientas de traducción de idiomas. De los tres tipos de IA que establece esta categorización, es el único que existe en la práctica hoy en día, por lo que cualquier otra forma de IA, en estos momentos, es meramente teórica.
- **IA general:** la AGI *(artificial general intelligence)* hace referencia a sistemas de IA con la capacidad de utilizar aprendizajes y habilidades previos para realizar nuevas tareas en un contexto diferente, sin la necesidad de un entrenamiento específico por parte de las personas. Tendría la capacidad de aprender y realizar cualquier tarea intelectual que un ser humano pudiera llevar a cabo. Esta tipología sigue siendo un concepto teórico y, en la actualidad, ningún sistema puede alcanzar este nivel de inteligencia.
- **Super IA:** esta categoría está integrada por aquellos sistemas de IA con capacidad para superar la inteligencia humana en la resolución de problemas, la creatividad y las habilidades generales. De este modo, desarrollaría emociones, deseos, necesidades y creencias propias y sería capaz de tomar sus propias decisiones y resolver sus propios problemas. Esta tipología también es, por el momento, estrictamente teórica.

A raíz de la aparición y popularización de la GenIA se ha abierto un extenso debate acerca de los límites éticos y sociales de la IA.

El segundo enfoque de clasificación ha sido establecido por Arend Hintze, profesor de Biología Integrada y Ciencias de la Computación de la University of Michigan, y distingue hasta cuatro tipos de IA, según su funcionalidad, es decir, en función del tipo de trabajo que pueden realizar (Naveen 2019; IBM Data and AI Team 2023):

- **Máquinas reactivas:** son el tipo de sistema de IA más antiguo y también el más básico. No tienen la capacidad de formar recuerdos, por lo que no pueden utilizar experiencias pasadas en las que basar la toma de decisiones actuales. Esto significa que carecen de la capacidad de aprender. Si bien es posible mejorar su capacidad para ejecutar mejor sus tareas específicas, no pueden aplicarse a otras situaciones. El ejemplo más paradigmático de máquina reactiva es la Deep Blue, una supercomputadora creada por IBM que fue capaz de vencer al ajedrez al maestro Garry Kasparov en 1996, tras analizar las piezas que había en el tablero y predecir los resultados probables de cada movimiento. Otro ejemplo más reciente es el motor de recomendaciones a usuarios de la plataforma Netflix, basadas en modelos que procesan conjuntos de datos recopilados del historial de visualización para ofrecerles el contenido que probablemente vayan a disfrutar.
- **Memoria limitada:** son aquellos sistemas que, además de contar con las capacidades de las máquinas puramente reactivas, también pueden aprender de datos históricos para tomar decisiones. Sin embargo, si bien pueden utilizar datos pasados durante un periodo de tiempo específico, no tienen la capacidad de retener dicha información en una biblioteca de experiencias pasadas para utilizarla durante un periodo de tiempo prolongado. Casi todas las aplicaciones de IA actuales –a saber, chatbots y asistentes virtuales– son de memoria limitada. Incluso los vehículos autónomos utilizan este tipo de IA para comprender el mundo que los rodea en tiempo real y tomar decisiones informadas sobre cuándo aumentar la velocidad, frenar, girar, etc.

Si bien los dos tipos anteriores de IA han sido ya desarrollados y tienen una relevante presencia en nuestro día a día, los dos siguientes por el momento solo existen como conceptos o trabajos en progreso.

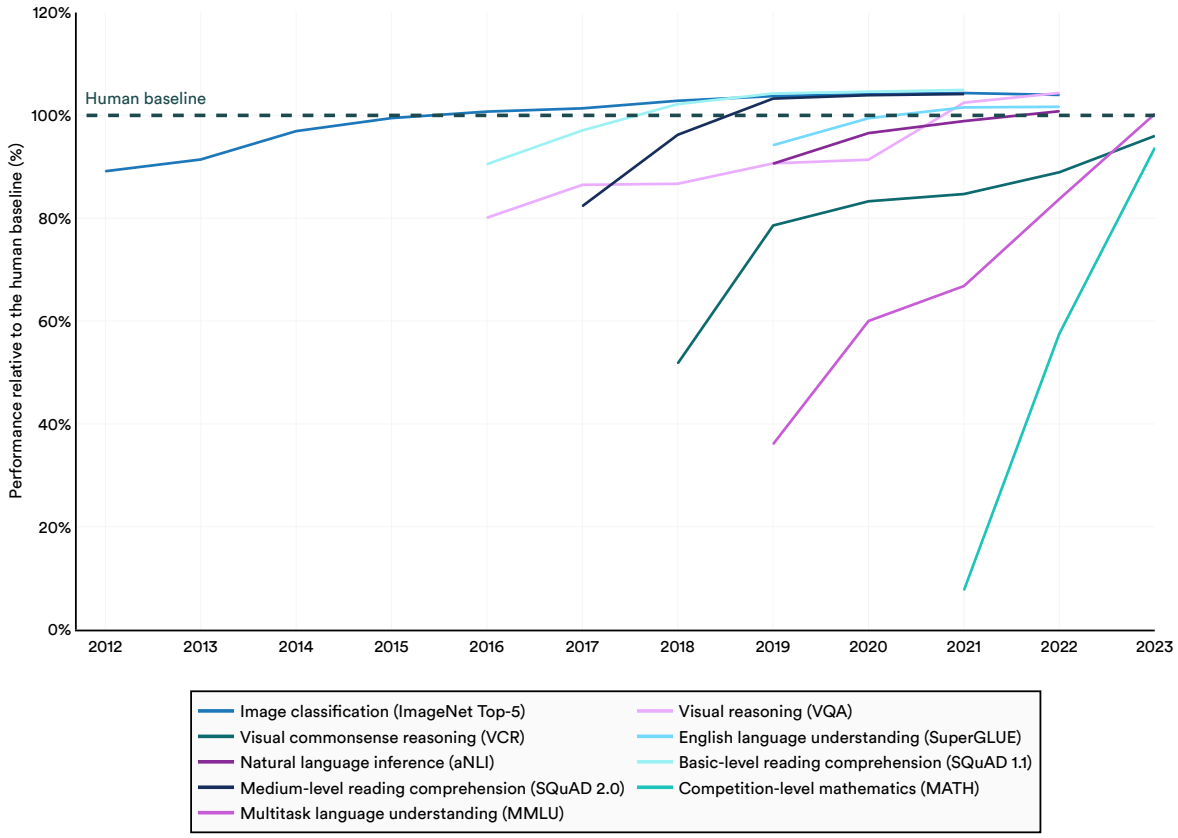
- **Teoría de la mente:** esta tipología consiste en un sistema con la capacidad de comprender mejor las entidades con las que interactúa al discernir sus necesidades, emociones, creencias y procesos de pensamiento. En principio, esto permitiría a la IA simular reacciones similares a las humanas. De este modo, al inferir los motivos y el razonamiento humanos, personalizaría sus interacciones con los individuos en función de sus necesidades e intenciones emocionales. También sería capaz de comprender y contextualizar obras de arte y ensayos, algo que las herramientas de GenIA actuales no pueden hacer. Esta IA de las emociones es un proyecto en desarrollo que los investigadores esperan que llegue a tener la capacidad de analizar voces, imágenes y otros tipos de datos para reconocer, simular, monitorear y responder de forma adecuada a las personas a nivel emocional.
- **Autoconciencia:** esta categoría representa la etapa final del desarrollo de la IA y, en la actualidad, solo existe de forma hipotética. En ella, estos sistemas habrían evolucionado hasta tal similitud con el cerebro humano que habrían desarrollado autoconciencia, esto es, la capacidad de formar representaciones sobre sí mismos. En caso de lograrse, se generarían sistemas con competencia para comprender sus propias condiciones y rasgos internos, junto con las emociones y los pensamientos humanos, además de tener sus propias emociones, necesidades y creencias.

Tal como se ha indicado en la introducción de este cuaderno, la IA lleva desarrollándose décadas. Ya en 1950 el matemático inglés Alan Turing publicó en la revista *Mind* un artículo titulado "Computing Machinery and Intelligence" que abrió las puertas al campo que posteriormente se llamaría *IA.* Posteriormente, en 1956 se llevó a cabo el Proyecto de Investigación de Verano de Dartmouth sobre Inteligencia Artificial, una conferencia con destacados investigadores de diversos campos para un debate abierto sobre esta tecnología, que es considerada el nacimiento de esta como campo de investigación científica y donde se acuñó por primera vez el término *inteligencia artificial* (McCarthy *et al.* 1955). En esa misma conferencia se presentó Logic Theorist, un programa diseñado para imitar las habilidades de resolución de problemas de un ser humano, considerado por muchos el primer programa de IA.

Desde entonces, han ido desarrollándose múltiples sistemas de IA, cada vez más complejos y aplicables a una mayor cantidad de ámbitos y funcionalidades, lo que ha permitido que estos hayan ido integrándose en numerosos sectores: salud, finanzas, transporte, educación, manufactura, etc. En la actualidad, indicadores recientes señalan que las capacidades de la IA han alcanzado niveles de rendimiento que superan las capacidades humanas en toda una serie de tareas (véase la **Figura 1**). A pesar de ello, por el momento todavía existen algunas áreas donde no las supera, en concreto, en tareas cognitivas complejas como el razonamiento visual, de sentido común o la resolución de problemas matemáticos de nivel avanzado.

Se han desarrollado sistemas de IA cada vez más complejos y aplicables a una mayor cantidad de ámbitos, funcionalidades, y sectores.

Figura 1. Puntos de referencia de rendimiento técnico del índice de IA frente al rendimiento humano



Fuente: Maslej et al. (2024).

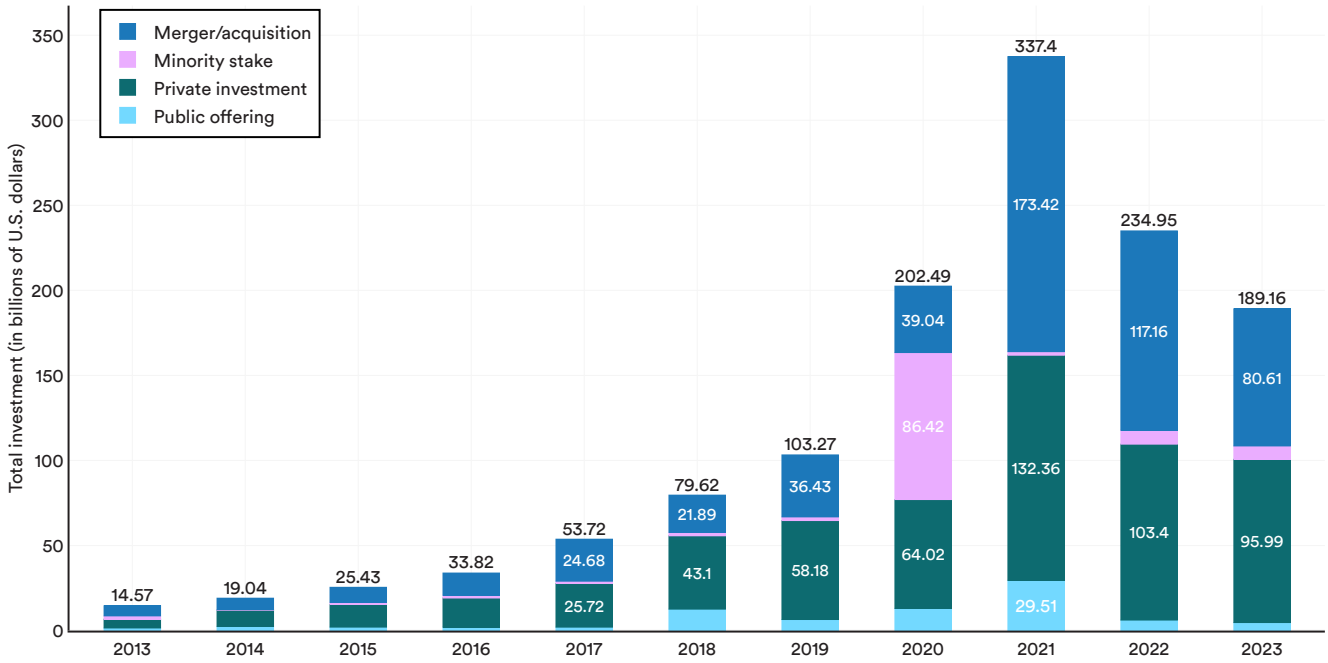
En cuanto a la opinión pública respecto de estos avances tecnológicos, si bien varía en gran medida según el país, el 54% de los encuestados a nivel mundial está de acuerdo en que la IA mejorará la eficiencia de sus tareas, el 39% cree que beneficiará su salud y el 37% piensa que mejorará su trabajo. Solo el 34% prevé que impulsará la economía y solo también el 32% cree que mejorará el mercado laboral. Es decir, la percepción pública reconoce sus ventajas tecnológicas, pero también manifiesta ciertas preocupaciones sobre sus posibles implicaciones económicas y laborales (Ipsos 2023).

Impacto en el mundo empresarial

El ámbito de las empresas no es ajeno a toda esta revolución de la IA. Desde EY (2024) destacan que la GenIA se ha convertido, en la actualidad, en una prioridad de inversión, lo cual, de hecho, ya llevan a cabo el 43% de las compañías consultadas al respecto. Esta información se corrobora desde la Stanford University (Maslej *et al.* 2024), que asegura que, en la última década, la inversión corporativa en esta tecnología se ha multiplicado por 13 a nivel global, con un pico de crecimiento en el 2021, cuando las inversiones fueron hasta 30 veces superiores respecto a las registradas en el 2013.

Parece, no obstante, que este crecimiento se ha visto frenado hace poco. Tal como podemos observar en la **Figura 2,** por primera vez en esta última década los datos reflejan una disminución de la inversión corporativa en IA de forma consecutiva en los años 2022 y 2023. Nestor Maslej, gerente de Investigación del Instituto Stanford para la Inteligencia Artificial Centrada en el Humano (HAI), considera que este escenario no representa una desaceleración, sino que, por el contrario, es indicativo de que, a medida que la tecnología de la IA madura, se vuelve cada vez más competitivo el lanzamiento de nuevas empresas de este ámbito. En este sentido, la reducción del número de nuevas compañías significa que la inversión se está concentrando: "Mayores cantidades de dinero van a menos jugadores" (Quach 2023).

Figura 2. Inversión corporativa en IA a nivel global en el periodo 2013-2023

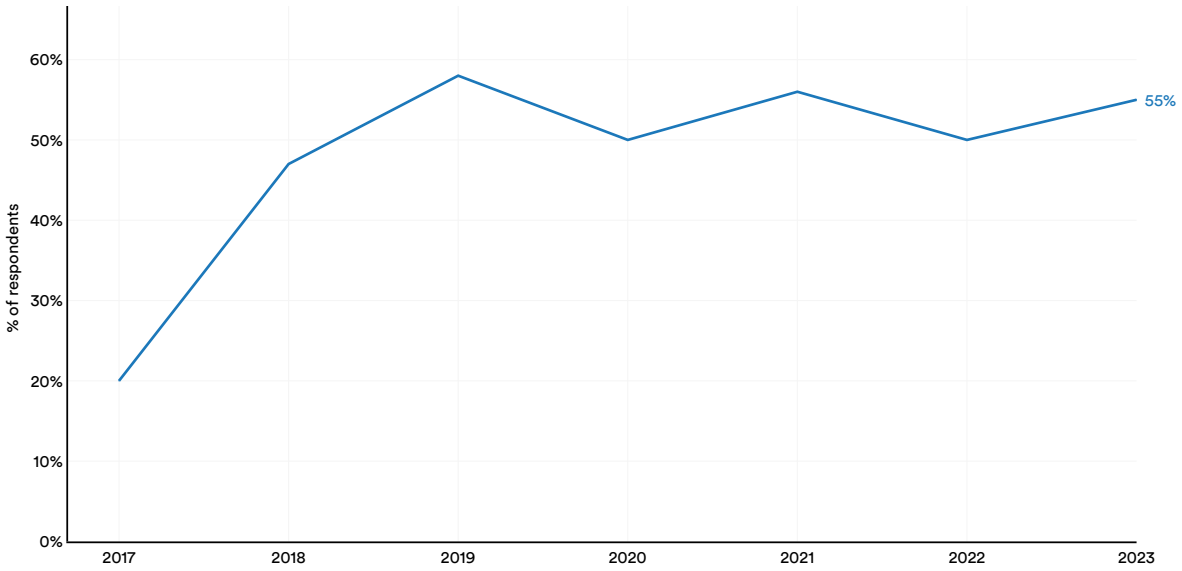


Fuente: Maslej et al. (2024).

Según una reciente encuesta global realizada por McKinsey (QuantumBlack, AI by McKinsey 2024) a 1363 representantes de toda la gama de regiones, industrias, tamaños de empresas, especialidades funcionales y permanencias, un 55% de las organizaciones , en el 2023, aseguró haber adoptado la IA en, al menos, un área del negocio, cifra que representa más del doble de los datos reportados en el 2017, cuando tan solo el 20% de las encuestadas afirmaron haberlo hecho. Tal como puede observarse en la **Figura 3,** en los últimos años se ha experimentado un crecimiento exponencial en el porcentaje de compañías que han integrado tales sistemas con el objetivo de fortalecer funciones empresariales de toda clase: desde la recopilación y el análisis de los datos hasta, incluso, operaciones más complejas como la automatización de procesos, la optimización de funciones comerciales, el servicio de atención al cliente o la gestión de riesgos.

En los últimos años se ha experimentado un crecimiento exponencial en el porcentaje de compañías que han integrado la IA para fortalecer funciones empresariales de toda clase.

Figura 3. Porcentaje de empresas encuestadas a nivel global que aseguran haber adoptado la IA en, al menos, una función en el periodo 2017-2023



Fuente: Maslej et al. (2024).

A continuación, se repasan las principales áreas y actividades en las que las empresas están implementando el uso de sistemas de IA.

a) Atención al cliente

La revolución en esta actividad empresarial se ha producido, principalmente, gracias al avance del llamado *procesamiento del lenguaje natural* (PLN), una tecnología de ML que dota a los sistemas de IA de la capacidad para interpretar, manipular y comprender el lenguaje humano. A través de los *softwares* de PLN es posible procesar de forma automática los datos recogidos, analizar la intención o el sentimiento de los mensajes y responder en tiempo real a la comunicación humana, mejorando de este modo la experiencia del usuario.

En este sentido, el área de atención al cliente es la que se ha visto más favorecida por las innovaciones en IA y en la que más compañías están implementando esta tecnología. Según una encuesta realizada por Forbes Advisor (Haan y Watts 2023), el 73% de las compañías que aseguraron estar utilizando sistemas de IA en su negocio se refirió al uso de chatbots con tecnología de IA para mensajería instantánea; un 61%, para la optimización de correos electrónicos; y un 55% aseguró estar utilizando servicios personalizados, como las recomendaciones de productos. Del mismo modo, se están produciendo avances en el manejo de llamadas telefónicas a través de la IA, por lo que el 36% de las empresas encuestadas aseguraron estar usándola o planear hacerlo en este ámbito, y un 49%, en la optimización de los mensajes de texto.

En resumen, hay una tendencia cada vez más generalizada a integrar la IA en los distintos canales de interacción con clientes, convirtiendo la experiencia general de estos en un área progresivamente más eficiente y personalizada. Según un estudio realizado por el IBM Institute for Business Value (2024), más de cuatro de cada cinco empresas (84%) esperan utilizar asistentes de GenIA con los clientes para el 2025, frente al 42% que aseguró estar haciéndolo ya en el 2023.

b) Optimización de los procesos

La incorporación de la IA en las compañías también se está enfocando en el objetivo de adquirir una mayor agilidad y productividad. Según la encuesta mencionada de Forbes Advisor (Haan y Watts 2023), está siendo utilizada –o se prevé su utilización– en diversos aspectos de la gestión empresarial. Por un lado, un 53% de las organizaciones encuestadas aseguró aplicarla o querer aplicarla para mejorar los procesos de producción, ámbito en el que destaca, en particular, la gestión de la cadena de suministro (30%). Gracias a los sistemas de IA se pueden obtener grandes mejoras tanto en la planificación de la demanda y la gestión de inventarios, a través de predicciones algorítmicas, como en la logística y la distribución, mediante rutas de entrega más eficientes y la reducción del tiempo y los costes de transporte.

Por otro lado, un 51% de las empresas encuestadas (Haan y Watts 2023) aseguró estar utilizando la IA para la automatización de tareas repetitivas y rutinarias, y un 40%, para la recopilación y el análisis de datos. Respecto de esta segunda

acción, la IA permite tanto analizar grandes volúmenes de datos de manera rápida y precisa como prever cuestiones como tendencias futuras, demanda de productos, comportamiento de los clientes o posibles riesgos financieros. Por otra parte, la implementación de herramientas de recopilación de datos también requiere de mecanismos con los que poder administrar y almacenar correctamente tal información.

En este sentido, la IA también juega un papel importante a la hora transformar los datos recogidos en información ordenada, clasificada, de utilidad y correctamente interpretada. Por último, otro ámbito en el que un alto porcentaje de las compañías (46%) afirman estar implementando sistemas de IA es la agilización de las comunicaciones, planes, presentaciones e informes internos, lo cual facilita una transferencia de datos más fluida y ágil entre departamentos.

c) Ciberseguridad y gestión del fraude

En este ámbito, el citado estudio de Forbes Advisor (Haan y Watts 2023) señala que más de la mitad (51%) de las empresas utilizan o se plantean utilizar en un futuro utilizar sistemas de IA para minimizar los riesgos de seguridad, ya que estos pueden analizar grandes cantidades de datos en tiempo real, identificar patrones y detectar anomalías significativas de posibles actividades fraudulentas antes de que se materialicen. Además, la esta tecnología aprende con datos históricos, por lo que puede incluso ajustar sus reglas para detener amenazas no vistas anteriormente. Este aprendizaje continuo permite a las compañías mantenerse un paso por delante de los atacantes, optimizando la seguridad de sus operaciones y protegiendo mejor los datos sensibles.

En resumen, la adopción de sistemas de IA para la seguridad empresarial es una tendencia creciente que permite a las organizaciones analizar datos en tiempo real, detectar y prevenir actividades fraudulentas y aprender de datos históricos para adaptarse a nuevas amenazas. Todo ello se traduce en sistemas de seguridad cada vez más robustos y muy demandados debido a las crecientes necesidades de protección en un entorno digital cada vez más complejo.

d) *Marketing* y ventas

La implementación de la IA en este ámbito se basa, principalmente, en el uso de herramientas y programas de análisis de datos. A través de estos recursos, las compañías pueden procesar grandes volúmenes de información de un modo eficiente, identificando patrones y tendencias esenciales para la toma de decisiones estratégicas. Asimismo, la capacidad de hacerlo en tiempo real les permite adaptarse con rapidez a los cambios del mercado y a las necesidades de los consumidores.

La principal ventaja de su aplicación este ámbito para las empresas es la capacidad de generar un *marketing más personalizado.* En la encuesta de Forbes Advisor (Haan y Watts 2023), un 55% de las compañías aseguró utilizar la IA para ofrecer servicios personalizados; un 46%, para publicidad personalizada; y un 42%, para la generación de contenidos web. Esta tecnología tiene la capacidad de analizar grandes volúmenes de datos sobre el comportamiento y las preferencias de los clientes, lo que permite segmentar mercados de manera más efectiva, identificando patrones y comportamientos en los consumidores, lo cual, a su vez, facilita la creación de grupos de clientes con características similares. De este modo, a través de esta información, las compañías pueden crear campañas de *marketing más personalizadas,* dirigidas específicamente a segmentos de mercado que tienen una mayor probabilidad de estar interesados en ciertos bienes o servicios, así como optimizar el *timing* de dichas campañas, asegurando que los mensajes y las ofertas lleguen a las personas correctas en el momento adecuado.

Esta tecnología permite analizar grandes volúmenes de datos, predecir tendencias del mercado, personalizar la experiencia de los clientes o mejorar las cadenas de suministro.

Por otro lado, a través de estos sistemas de IA, las empresas también pueden predecir el comportamiento de sus clientes. Con modelos predictivos avanzados, es posible analizar patrones históricos de comportamiento, preferencias de compra y datos demográficos con los que anticiparse a las necesidades y los deseos de sus clientes. Estos modelos emplean técnicas como el ML y el análisis de *big data* para identificar correlaciones y tendencias que no serían evidentes a simple vista, lo que les permite diseñar estrategias efectivas. Es el caso, por ejemplo, de la creación de programas de fidelización: al comprender mejor lo que motiva a los clientes y lo que los hace retornar, las compañías pueden diseñar mejores acciones acordes a los intereses y preferencias de estos, logrando aumentar su retención y lealtad.

Asimismo, resulta frecuente que toda esta labor de recopilación de datos se vea complementada por el uso de GenIA. Gracias a la información adquirida, es posible producir contenidos diversos como artículos, imágenes, videos y otros materiales multimedia, de un modo personalizado. Además la capacidad de aprender de los datos recopilados que brin-

da facilita que la producción de contenidos se vaya adaptando en tiempo real, a los cambios en las preferencias y necesidades de los consumidores.

e) Finanzas y contabilidad

En cuanto a las actividades de carácter financiero y contable, los sistemas de IA también permiten a las empresas la capacidad de procesar grandes volúmenes de datos, identificar patrones y prever posibles escenarios con una gran precisión. Por lo tanto, también facilita a los responsables de tomar decisiones estratégicas hacerlo basándose en predicciones confiables y en análisis de tendencias del mercado en tiempo real. En este sentido, la implementación de estos sistemas en las áreas de finanzas y contabilidad permite a las compañías identificar de un modo temprano tanto riesgos como oportunidades, optimizando de este modo la gestión del capital y la planificación financiera. Según Forbes Advisor (Haan y Watts 2023), un 30% de las empresas encuestadas aseguró haber incorporado esta tecnología en sus actividades de contabilidad.

Además, la IA puede automatizar tareas rutinarias como la conciliación de cuentas, mediante la comparación, a través de algoritmos de ML, de las transacciones bancarias con los registros contables, pudiendo detectar discrepancias. Asimismo, hace posible recopilar datos de múltiples fuentes, aplicar normas contables y generar informes detallados en cuestión de minutos, lo que no solo acelera el proceso de reporte financiero, sino que también garantiza una mayor exactitud y evita el riesgo de posibles errores humanos.

Del mismo modo, estos sistemas también pueden ayudar en el cumplimiento de las regulaciones financieras. Gracias a su fácil adaptación a los cambios en las normativas, garantizan que las operaciones de tesorería cumplan con los requisitos legales vigentes, lo cual, además, reduce el riesgo de recibir sanciones y multas por incumplimiento. Para ello, los algoritmos de IA revisan automáticamente transacciones y procesos, asegurándose de que se adhieran a las normativas específicas y alertando de cualquier irregularidad detectada a los responsables.

f) Recursos humanos

Según un estudio realizado en Estados Unidos por la Society for Human Resource Management (SHRM 2024) –la organización profesional de recursos humanos (RR. HH.) más grande del mundo–, un 28% de las empresas aseguró estar usando en 2023 IA para apoyar actividades en este ámbito. Las que no lo hacen aluden, principalmente, a la falta de conocimiento respecto a qué herramientas se adaptarían mejor a sus necesidades (42%), la falta de recursos (41%) y la ausencia de "toque humano" que implica esta tecnología (40%). Por el contrario, las que la han incorporado lo han hecho, sobre todo, en actividades relacionadas con el reclutamiento, las entrevistas y la contratación (64%), el aprendizaje y desarrollo (43%) y la gestión del desempeño (25%).

En lo que respecta a las actividades de reclutamiento, entrevistas o contratación, prácticamente la totalidad (88%) de las organizaciones que incorporan la IA aseguran hacerlo para ahorrar tiempo o aumentar su eficiencia, lo cual les permite priorizar y atender tareas que requieren exclusivamente inteligencia humana. Por otro lado, casi dos de cada tres la utilizan en este ámbito para ayudar a generar sus descripciones de trabajo cuando publican ofertas de empleo, con el fin de personalizarlas o dirigirlas a grupos específicos. Asimismo, un 33% asegura que la diversidad de las contrataciones de su organización ha mejorado gracias al uso de esta tecnología, al posibilitarles acceder a grupos de talentos subrepresentados a los que antes no llegaban.

En cuanto a su implementación para el aprendizaje y desarrollo, casi la mitad de las organizaciones que la han adoptado a tal efecto aseguran utilizarla para recomendar o crear oportunidades personalizadas de formación y desarrollo para sus empleados, logrando programas de aprendizaje y desarrollo más efectivos (51%) y un mayor compromiso por parte de aquellos (44%). En este sentido, la IA puede ayudar a recomendar módulos de formación personalizados para, por ejemplo, la movilidad profesional. Al analizar los datos de cada trabajador, como sus habilidades y preferencias, puede adaptar su formación a sus objetivos personales, así como apoyar a los gerentes de RR. HH. a la hora de identificar talentos ocultos o a empleados listos para un ascenso (IBM Consulting 2023).

Sin desmerecer el innegable potencial que posee la IA para generar beneficios significativos, resulta fundamental no pasar por alto los desafíos asociados a su implementación.

Por último, las empresas que la aplican para la gestión del desempeño lo hacen con el fin de contribuir en la facilitación de las conversaciones sobre el rendimiento y los siguientes pasos. En concreto, más de la mitad la utiliza para ayudar

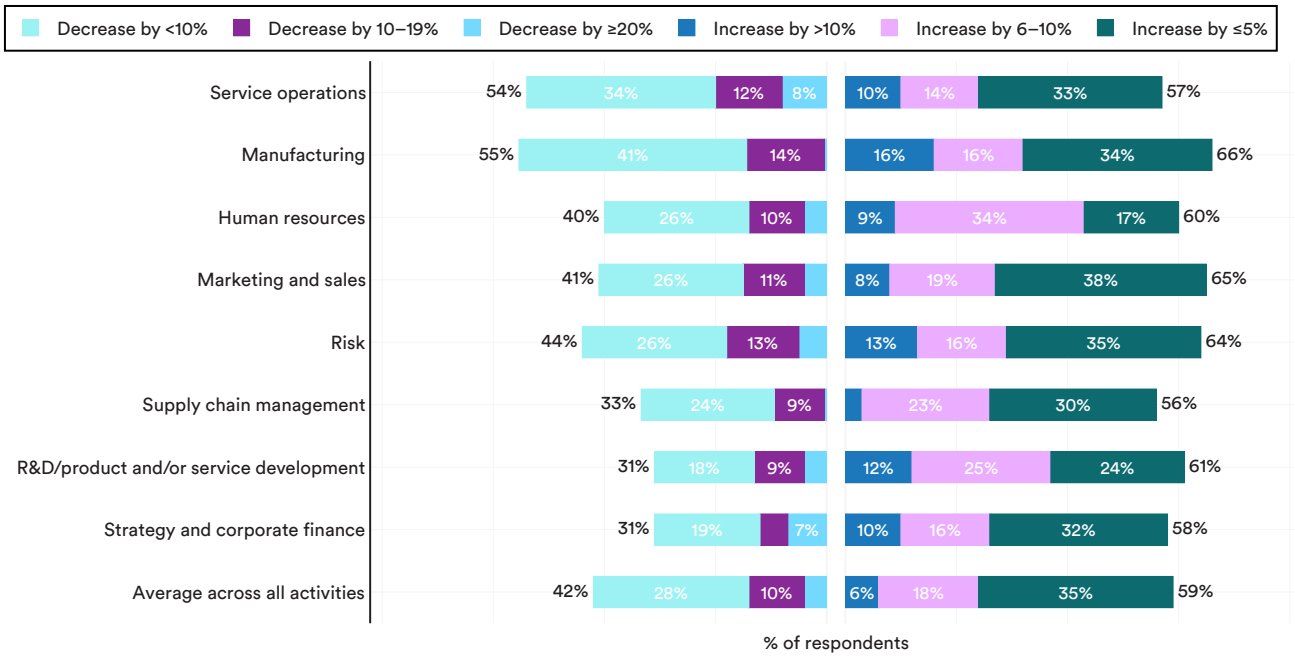
a sus responsables de personal a brindar comentarios más completos o procesables para sus empleados, y el 46%, para facilitar establézcanla definición de objetivos alcanzables y personalizado y la monitorización del desempeño.

Riesgos éticos asociados en el ámbito empresarial

El auge de la IA, tal como se ha visto, ha generado nuevas oportunidades a nivel global y ha demostrado su potencial para mejorar nuestra calidad de vida, al poder facilitar diagnósticos de salud, fomentar y ampliar las redes de conexión humanas a través de las redes sociales o al aumentar la eficiencia de nuestras actividades mediante la automatización de tareas. Sin embargo, al mismo tiempo, esta realidad ha puesto sobre la mesa toda una serie de dilemas éticos. Por tanto, sin desmerecer el innegable potencial que posee para generar beneficios significativos, resulta fundamental no pasar por alto los desafíos asociados a su implementación.

El ámbito empresarial no es una excepción en este sentido. Tal como también se ha visto, esta tecnología permite a las compañías analizar grandes volúmenes de datos, predecir tendencias del mercado, personalizar la experiencia de sus clientes y mejorar las cadenas de suministro, entre otras muchas aportaciones (véase la **Figura 4**).

Figura 4. Reducción de costes y aumento de ingresos gracias a la adopción de IA, por función, en el 2022



Fuente: Maslej et al. (2024).

Sin embargo, su implementación también plantea a las organizaciones una serie de desafíos a nivel ético que requieren atención y análisis por su parte, además de por la comunidad tecnológica, los legisladores y la sociedad en general. Es fundamental poder garantizar que el uso y el desarrollo de la IA por parte de las compañías se lleven a cabo de manera responsable y equitativa, con respeto a los derechos humanos y promoción del bienestar social.

Existe la preocupación generalizada de cómo los sistemas de IA pueden perpetuar y hasta amplificar sesgos y prejuicios existentes en los datos con los que son entrenados.

La lista de temas que plantean alguna consideración ética, tal como se anticipaba en la introducción de este cuaderno, va ampliándose a medida que se desarrolla esta tecnología. En concreto, en la actualidad hay tres aspectos que están adquiriendo una especial relevancia, por lo que nos detendremos en ellos a continuación: los sesgos introducidos en algoritmos y los problemas de discriminación que implican, la seguridad en el almacenamiento y la gestión de los datos, y la necesidad de una transparencia que aporte fiabilidad en el uso de la IA.

a) Sesgo y discriminación

Existe la preocupación generalizada de cómo los sistemas de IA pueden perpetuar y hasta amplificar sesgos y prejuicios existentes en los datos con los que son entrenados. Tal como declara Michael Sandel, profesor de la Facultad de Derecho de la Harvard University, gran parte del atractivo de la posibilidad de tomar decisiones de forma algorítmica es que parece ofrecer una forma objetiva con la que poder superar la subjetividad y los juicios de valor humanos (Pazzanese 2020). Sin embargo, el problema radica en que muchos de estos algoritmos replican e incorporan los prejuicios que ya existen en nuestra sociedad, en la medida en que son entrenados con datos e informaciones impregnados de ellos. De este modo, existe un riesgo alto de que esto provoque decisiones automatizadas que discriminen injustamente a ciertos grupos sociales, contribuyendo a la desigualdad y la injusticia. Karen Mills, profesora del Harvard Business School, alude, por ejemplo, al caso de las decisiones basadas en algoritmos en el sector bancario, donde, "en la medida en que las máquinas aprenden de los conjuntos de datos que reciben, las probabilidades son bastante altas de que puedan replicar muchas de las fallas pasadas de la industria bancaria que resultaron en un trato sistemático y desigual a los afroamericanos y otros consumidores marginados" (Pazzanese 2020).

Un ámbito empresarial especialmente vulnerable a este aspecto es el de la contratación de personal, donde el uso de la IA ha crecido de forma significativa en los últimos años gracias a su potencial para optimizar los procesos de selección, reducir el tiempo y los costes asociados y mejorar la eficiencia en la identificación de candidatos cualificados. Sin embargo, esto también ha comportado toda una serie de riesgos con graves implicaciones para las organizaciones y los candidatos. Si un algoritmo se entrena con datos históricos de contratación en los que ciertos grupos demográficos se han visto favorecidos por encima de otros, replicará dichos sesgos, discriminando a candidatos por razones de género, raza, edad u otros factores. Es el caso, por ejemplo, de las personas con diversidad funcional, quienes han sido históricamente objeto de marginación y han contado con menores oportunidades de acceso al poder y a los recursos. En este sentido, cabe suponer que dichos patrones de marginación están inscritos en los datos que dan forma a los sistemas de esta tecnología y son integrados en su lógica interna.

Un caso paradigmático de ello, por ejemplo, tuvo lugar en Amazon, empresa en la que la automatización ha resultado clave para el dominio del comercio electrónico. Según hicieron saber empleados anónimos de esta compañía a la agencia de noticias Reuters (Dastin 2018), Amazon puso en marcha en el 2014 programas informáticos de IA para revisar los currículums de los solicitantes de empleo con el objetivo de mecanizar la búsqueda de los mejores talentos. Su labor consistía, principalmente, en otorgar a los candidatos puntuaciones. Sin embargo, en el 2015, la empresa se dio cuenta de que su nuevo sistema no clasificaba a los candidatos para puestos de desarrollador de *software* y otros puestos técnicos de forma neutral en lo relativo al género. El motivo era que dichos modelos informáticos habían sido entrenados para examinar a los solicitantes a partir de la observación de patrones en los currículums presentados a la compañía durante un periodo de 10 años, y dado que la mayoría procedían de hombres, el sistema asumió que los candidatos masculinos eran preferibles, penalizando la presencia de palabras como *femenino* o degradando aquellos currículums en los que aparecía el nombre de universidades exclusivas para mujeres. Finalmente, Amazon disolvió el proyecto en el 2017 (Dastin 2018).

Por este motivo, en la actualidad predomina la creencia de que un sistema de IA, en principio, no debería procesar información relacionada con datos sensibles como el origen racial

o étnico, las opiniones políticas, la religión, las creencias o la orientación sexual, para evitar, precisamente, que esto conduzca a un tratamiento arbitrario que derive en discriminación o sesgo por parte de estos sistemas.

b) Privacidad y la seguridad de los datos

La IA también plantea preocupaciones significativas en cuanto a la privacidad y la seguridad de los datos, ya que el uso de sus algoritmos implica, necesariamente, la recopilación y el análisis de grandes volúmenes de información que incluye desde datos personales hasta patrones de comportamiento y preferencias individuales. La capacidad de esta tecnología para recopilar, analizar y procesar estos datos genera inquietudes sobre cómo se protege y se utiliza dicha información, dado que esta recopilación masiva de datos, por un lado, puede conducir a vulneraciones de la privacidad si no se gestiona adecuadamente y, por otro, despierta la preocupación de que aquellos puedan llegar a ser utilizados sin el consentimiento explícito de las personas.

Así pues, la seguridad de los datos representa hoy en día una de las principales preocupaciones en este ámbito. Según una consulta realizada por McKinsey (QuantumBlack, AI by McKinsey 2024) a representantes de 1363 empresas de todo el mundo respecto al uso de sistemas de IA, más de la mitad de ellas concentró sus mayores preocupaciones tanto en lo relativo a la ciberseguridad y la protección de los datos (51%) como en cuestiones relacionadas con la posible vulneración de la privacidad personal (43%).

En lo que respecta a las filtraciones de datos, estas pueden ocurrir debido a vulnerabilidades en los sistemas de seguridad que protegen la información recopilada por los sistemas de IA. En la medida en que estos sistemas recopilan y almacenan grandes volúmenes de datos personales en sus bases de datos, se convierten en objetivos atractivos para ciberdelincuentes. De este modo, una filtración de datos puede revelar información sensible sobre los individuos, como detalles financieros, historiales médicos o actividades en línea, exponiéndola al riesgo ser utilizada en robos de identidad, fraudes financieros y otros tipos de explotación.

En cuanto al uso indebido de la información personal por parte de las empresas o entidades que desarrollan y emplean algoritmos de IA, existen diversos riesgos significativos. En algunos casos, los datos recopilados pueden ser utilizados para fines no previstos originalmente, como la creación de perfiles detallados de los usuarios para la publicidad dirigida, la manipulación de opiniones o, incluso, la discriminación en la toma de decisiones automatizadas. En este sentido, si los datos no se manejan de manera ética y transparente, los individuos pueden perder el control sobre su propia información y afrontar consecuencias negativas en términos de privacidad y autonomía.

Un ejemplo de ello es el caso del escándalo de datos de Facebook-Cambridge Analytica, cuando, en la década del 2010, se descubrió que la consultora británica Cambridge Analytica recopiló datos de millones de usuarios de Facebook sin su consentimiento, principalmente para utilizarlos con un fin de propaganda política. Los datos se obtuvieron por medio de una aplicación llamada *This Is Your Digital Life,* la cual consistía en una serie de preguntas para elaborar perfiles psicológicos de usuarios y recabó los datos personales de los contactos de sus usuarios mediante la plataforma Open Graph, de Facebook. La aplicación logró hacerse con datos de hasta 87 millones de perfiles de esta red social, y Cambridge Analytica los utilizó para proporcionar asistencia analítica a las campañas de Ted Cruz y Donald Trump para las elecciones presidenciales del 2016 (*BBC Mundo* 2018).

La posibilidad de que estos sistemas recopilen datos de las personas sin su consentimiento explícito o, incluso, sin que se den cuenta plantea serias inquietudes sobre la privacidad y la protección de aquellos, ya que los usuarios pueden no estar informados sobre qué tipo de información se está recolectando, cómo se está utilizando y con quién se está compartiendo. Esta preocupación se intensifica cuando consideramos que los datos recolectados pueden incluir información sensible como hábitos de navegación, ubicaciones geográficas, patrones de comportamiento y hasta detalles íntimos de la vida personal y profesional. Por ello, resulta fundamental que las entidades que utilizan y desarrollan estas tecnologías sean transparentes e implementen políticas claras y estrictas sobre la recopilación y el manejo de datos, asegurando que los usuarios puedan ser plenamente conscientes y otorguen su consentimiento explícito antes de que cualquier información personal sea recopilada.

c) Transparencia y explicabilidad

Otro aspecto importante que se debe considerar en el uso de sistemas de IA es la transparencia y la explicabilidad de sus algoritmos. La explicabilidad aquí se refiere a la capacidad de los sistemas para proporcionar descripciones comprensibles y significativas acerca de cómo se alcanzan las decisiones. En muchos casos, estos algoritmos operan como cajas negras, lo que significa que sus procesos internos son opacos y difíciles de comprender, lo cual representa un obstáculo significativo a la hora de comprender cómo y por qué un sistema de IA ha llegado a una determinada decisión o predicción. Esta falta de transparencia provoca, de forma inevitable, una desconfianza por parte tanto de los usuarios –quienes no pueden verificar la validez o la justicia de las decisiones tomadas por esta tecnología– como de los desarrolladores –para quienes se complica la identificación de errores y sesgos dentro del sistema–. Si no existe la posibilidad de examinar y comprender cómo se llega a una conclusión, es difícil confiar plenamente en la tecnología.

Así pues, la naturaleza de caja negra de muchos algoritmos de IA implica que, aunque los resultados puedan ser realmente precisos y útiles, los métodos y criterios utilizados para llegar a ellos no sean fácilmente accesibles ni interpretables. Esto es problemático de forma especial, por ejemplo, en situaciones donde la toma de decisiones debe ser claramente justificable, como podría ser el caso de las evaluaciones de crédito, los diagnósticos médicos, las decisiones judiciales o la contratación laboral: sin una explicación clara acerca de cómo se tomó una decisión, será difícil evaluar su justicia, precisión y relevancia, existiendo en todo momento la desconfianza o el temor de estar perpetuando errores y sesgos preexistentes en los datos de entrenamiento.

Un ejemplo de ello es el algoritmo de fijación de precios de la empresa estadounidense Uber. Según ella, este algoritmo, conocido como *surge pricing* o "tarificación dinámica", ajusta de forma automática las tarifas en función de la oferta y la demanda en tiempo real (Uber, s. f.). Sin embargo, esta tecnología es propiedad de la compañía y está protegida como secreto comercial, lo que implica que no se revelan los detalles sobre cómo se calculan exactamente las tarifas, lo cual ha generado problemas significativos y desconfianza. En el 2023, ante la denuncia de las subidas de precios por parte de los usuarios, Dara Khosrowshahi, su CEO, atribuyó el encarecimiento al aumento de los precios a causa de la inflación y al incremento de los costes de tiempo y mano de obra (*Economic Times* 2023). Pero un informe de *Forbes* contradijo esta explicación, revelando que los precios de Uber en Estados Unidos habían aumentado cuatro veces la tasa de inflación del 2018 al 2022 (Sherman 2023). En este sentido, la falta de transparencia en la fórmula del algoritmo no permite a los usuarios comprender de forma plena por qué los precios aumentan de una forma tan drástica, lo que ha llevado a cuestionamientos sobre la equidad y la ética en la práctica de precios dinámicos de esta compañía.

Asimismo, la falta de explicabilidad también plantea desafíos significativos en términos de responsabilidad. Cuando la IA toma una decisión incorrecta o perjudicial, identificar quién es responsable se convierte en una tarea compleja: ¿el desarrollador del algoritmo, la empresa que lo implementó o el conjunto de datos utilizado para entrenar esta tecnología? Esta ambigüedad puede resultar en una falta de rendición de cuentas, donde ninguna de las partes involucradas asuma la responsabilidad total, lo que puede ser perjudicial para los afectados por las decisiones de esta tecnología.

La capacidad de esta tecnología para recopilar, analizar y procesar datos personales genera inquietudes sobre cómo se protege y se utiliza dicha información.

Propuestas reguladoras

El desarrollo de la IA ha tenido lugar sin regulaciones y legislaciones específicas al respecto, lo que ha provocado una implementación y un uso sin las directrices claras necesarias para garantizar la seguridad y protección de los derechos de

las personas. Sin embargo, ante esta ausencia de regulaciones gubernamentales y en respuesta a las preocupaciones crecientes al respecto, a lo largo de estos años grupos de expertos e investigadores en este ámbito, procedentes de disciplinas y regiones diversas, han elaborado diferentes orientaciones y recomendaciones con las que establecer unos principios que garanticen un desarrollo y un uso responsables de la IA.

Dada su trascendencia, se citan a continuación algunas de estas propuestas.

a) Recomendación sobre la ética de la inteligencia artificial del 2021 de la UNESCO

En noviembre del 2021, la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura elaboró la primera norma mundial sobre ética de la IA (UNESCO 2021), un marco adoptado por los 193 Estados miembros. Esta norma se centra en ofrecer orientaciones a los responsables políticos encargados de las legislaciones relativas a los sistemas de IA con el objeto de poder traducir toda una serie de principios éticos a la acción política. Presta atención, sobre todo, a cuestiones relativas a la protección de los derechos humanos y la dignidad; menciona principios fundamentales como la transparencia y la equidad; e insiste en la importancia de que se lleve a cabo la supervisión humana de esos sistemas.

b) Conceptos fundamentales para la IA generativa del Foro Económico Mundial (FEM)

El Global Future Council on Artificial Intelligence for Humanity del Foro Económico Mundial publicó, en el 2023, un documento en el que se establecieron una serie de pasos fundamentales en el desarrollo de la IA con los que poder garantizar su carácter inclusivo (Foro Económico Mundial 2023).

A través de unas normas claras y vinculantes y de unos de mecanismos de supervisión es posible proporcionar la estructura necesaria con la que garantizar una IA ética y responsable.

c) Principios de la OCDE sobre IA

La Recomendación sobre Inteligencia Artificial (OCDE 2024) representó el primer estándar intergubernamental sobre esta tecnología y fue adoptada por la reunión del Consejo de la OCDE el 22 de mayo del 2019. Su objetivo es fomentar la innovación y la confianza en esta tecnología mediante la promoción de una gestión responsable y, al mismo tiempo, garantizando el respeto de los derechos humanos y los valores democráticos. En junio del 2019, en la Cumbre de Osaka, los líderes del G20 acogieron los "Principios de IA", extraídos de la citada Recomendación. Este año (2024), ha sido revisada por el Consejo de la OCDE para actualizar la definición de *sistema de IA* e incorporar los avances tecnológicos y políticos que han tenido lugar desde entonces.

En esa Recomendación, la OCDE, por un lado, propone una comprensión común a partir del establecimiento de una definición conjunta de términos claves como *sistema de IA, ciclo de vida del sistema de IA* y actores de *IA;* por otro, recomienda la adopción de toda una serie de principios con los que poder garantizar una gestión responsable de esta tecnología, así como distintas formas de asegurar el desarrollo de políticas nacionales e internacionales conformes con dichos principios.

Los principios en cuestión son los siguientes:

- Crecimiento inclusivo, desarrollo sostenible y bienestar.
- Respeto por el Estado de derecho, los derechos humanos y los valores democráticos, incluidas la equidad y la privacidad.
- Transparencia y explicabilidad.
- Robustez, seguridad y protección.
- Rendición de cuentas.

d) Directrices éticas de la Comisión Europea para una IA confiable

Estas directrices se configuran como un marco de referencia, desarrollado por un grupo de expertos de alto nivel sobre esta tecnología constituido por la Comisión Europea (CE), en junio del 2018, que pretende promover una IA fiable. La fiabilidad dentro de este ámbito se entiende en estas directrices a partir de tres componentes que deben satisfacerse a lo largo de todo el ciclo de vida del sistema: por un lado, esta tecnología ha de ser lícita, es decir, debe cumplir todas las leyes y los reglamentos aplicables; por otro, tiene que ser ética, de modo que se garantice el respeto de los principios y valores éticos;

y, por último, ha de ser robusta, tanto desde el punto de vista técnico como social, puesto que los sistemas de IA, incluso si las intenciones son buenas, pueden provocar daños accidentales. En este sentido, estas directrices pretenden ofrecer orientaciones sobre el fomento y la garantía de una IA que cumpla con dichos componentes (Comisión Europea 2019).

e) Marco de ética de IA del IEEE

El Instituto de Ingenieros Eléctricos y Electrónicos (IEEE[1]) también estableció su propio marco (Morandín-Ahuerma 2023) para las consideraciones éticas en el diseño, desarrollo, despliegue y uso de los sistemas de IA, con el objetivo de garantizar que las tecnologías digitales ayuden a las personas en sus labores. En este sentido, basan su propuesta principalmente en ocho principios:

- Derechos humanos
- Bienestar
- Control de los datos
- Eficacia
- Transparencia
- Responsabilidad
- Conciencia del mal uso
- Competencia

Si bien todas las recomendaciones descritas brindan marcos de referencia de gran utilidad, no cabe duda de que no pueden sustituir a un marco legislativo formal y obligatorio. Tal como se ha venido observando, la dependencia exclusiva de estas propuestas no es suficiente para prevenir potenciales daños en la sociedad, ya que carecen de fuerza coercitiva y dependen, en gran medida, de la voluntad de los desarrolladores y de las empresas para su implementación. Además, la diversidad de enfoques existentes y la falta de una estandarización global provocan una adopción desigual y una aplicación inconsistente de estos principios. Por tanto, el abordaje de manera efectiva de estos desafíos evidencia la necesidad de que estos marcos éticos se vean complementados con esfuerzos legislativos y reguladores. Únicamente a través de normas claras y vinculantes, junto con la implementación de mecanismos de supervisión y cumplimiento, es posible proporcionar la estructura necesaria con la que garantizar una IA ética y responsable.

En este sentido, este año (2024) ha tenido lugar un hito especialmente significativo, al aprobarse, en el ámbito de la UE, la primera gran normativa relativa a esta cuestión, conocida como *Ley de IA*. Asimismo, se espera que, a raíz de esta primera norma, surjan otras regulaciones equivalentes en el resto del mundo, ya que, en la medida en que las empresas tecnológicas globales operan en Europa, estas se van a ver obligadas a adaptar sus prácticas a dichas directrices. De este modo, con la Ley de IA se aspira a lograr una estandarización global en torno a dicha tecnología y se impulsa que aquellos países que no adopten regulaciones similares perciban el riesgo de quedarse atrás en cuanto a la seguridad, la ética y la responsabilidad en el uso de la IA y, en consecuencia, vean afectada su competitividad en el mercado global. En el siguiente apartado, se aborda con más detalle la mencionada ley.

[1] Es la mayor organización técnica profesional del mundo, con más de 420.000 miembros en más de 160 países, que se dedican al avance en la innovación tecnológica y a la excelencia en beneficio de la humanidad. Más adelante se aborda con más detalle en este cuaderno.

La Ley de Inteligencia Artificial de la Unión Europea

El Reglamento 2024/1689 de la Unión Europea por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento 2024), también llamado *Ley de IA de la UE,* es el primer marco jurídico integral sobre esta tecnología en todo el mundo y, pese a que ha sido desarrollado por la UE, ha nacido con la idea de fomentar una IA fiable tanto en el ámbito europeo como fuera de él. Es el resultado de toda una serie de esfuerzos iniciados en el 2018 con la primera Estrategia Europea de IA que, finalmente, han dado lugar a una normativa propuesta por la CE en el 2021 y aceptada en el 2024 por el Parlamento y el Consejo europeos. La Ley de IA forma parte de un paquete más amplio de medidas políticas de la UE destinadas a apoyar y reforzar el desarrollo, la adopción, la inversión y la innovación en cuanto a esta tecnología en toda la UE y garantizar que esta no atente contra la seguridad y los derechos fundamentales de las personas y las empresas. Entre estas medidas también se encuentran el paquete de innovación en materia de IA y el Plan Coordinado sobre Inteligencia Artificial.

El paquete de innovación mencionado consiste en un conjunto de medidas destinadas a apoyar a las compañías emergentes y a las pymes europeas en el desarrollo de una IA que respete los valores y las normas de la UE. Por su parte, el Plan Coordinado pretende acelerar la inversión en esta tecnología, implementar estrategias y programas de IA y alinear las políticas relativas a esta cuestión para evitar la fragmentación dentro del ámbito europeo.

Por su parte, la Ley de IA ha nacido con el objetivo de establecer un marco normativo de referencia, por lo que, en primer lugar, establece una definición uniforme de la IA:

> A los efectos del presente Reglamento se entenderá por "sistema de IA" un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales. (Reglamento 2024).

Esta ley quiere garantizar que aquellos sistemas que queden englobados en dicha definición y vayan a utilizarse en el ámbito de la UE sean seguros, transparentes, trazables, no discriminatorios y respetuosos con el medioambiente. Para ello, su contenido se enfoca de forma específica en identificar los riesgos derivados de la IA, estableciendo una jerarquía entre ellos y promoviendo distintas acciones y obligaciones, en función del nivel de riesgo asociado. En este sentido, la Ley de IA clasifica los sistemas en cuatro categorías de riesgo:

1. Riesgo inaceptable: en este nivel se enmarcan los sistemas que se consideran una amenaza para las personas y deben ser prohibidos. Se incluyen:

- Los que provocan una manipulación cognitiva y distorsión del comportamiento de personas o grupos vulnerables específicos –por su edad, discapacidad o situación social o económica–, pudiendo causar daños físicos o psicológicos tanto a ellos mismos como a otras personas.
- Los que establecen puntuaciones sociales, clasificaciones de personas en función de su comportamiento, estatus socioeconómico o características personales.
- Los de identificación biométrica en tiempo real y a distancia, como el reconocimiento facial. No obstante, existen excepciones, como los sistemas de identificación biométrica a distancia ***a posteriori,*** en los que la identificación se produce tras un retraso significativo. Se permiten para perseguir delitos graves y solo con una aprobación judicial previa.

2. Riesgo alto: aquellos sistemas que afectan de forma negativa a la seguridad o a los derechos fundamentales. Se dividen en dos categorías:

- Los que se utilicen en productos sujetos a la legislación de la UE sobre la seguridad de estos. Incluyen juguetes, aviación, automóviles, dispositivos médicos y ascensores.
- Los pertenecientes a los siguientes ocho ámbitos específicos cuya acción sea relevante en una decisión con posible riesgo sobre la salud, la seguridad o los derechos fundamentales:
  - Identificación biométrica y categorización de personas físicas sin su participación activa.
  - Gestión y explotación de infraestructuras críticas.
  - Educación y formación profesional.
  - Empleo, gestión de trabajadores y acceso al autoempleo.
  - Acceso a y disfrute de servicios privados esenciales y servicios y prestaciones públicas.
  - Aplicación de la ley y actividades de fuerzas y cuerpos de seguridad.
  - Gestión de la migración, el asilo y el control de fronteras.
  - Asistencia en la interpretación jurídica y aplicación de la ley.

La ley establece que, de forma obligatoria, todos los sistemas de IA de alto riesgo deberán ser evaluados antes de su comercialización y a lo largo de su ciclo de vida. Asimismo, para evitar resultados perjudiciales, se decreta que tales supervisiones tendrán que ser realizadas por personas y no a través de la automatización. Además, se establece que los ciudadanos tienen derecho a presentar reclamaciones sobre estos sistemas de IA de riesgo alto ante autoridades nacionales específicas.

3. Riesgo limitado: esta categoría se refiere a los posibles riesgos asociados con la falta de transparencia en el uso de la IA. En este sentido, se introducen una serie de obligaciones específicas para garantizar que los seres humanos estén informados de que están interactuando con esa tecnología, por ejemplo, cuando se utilizan sistemas como los chatbots. Asimismo, se establece la obligación de que el contenido generado por la IA sea identificable, debiendo etiquetarse claramente como tal. Así, por ejemplo, la GenIA , como ChatGPT, no se considera de alto riesgo siempre que cumpla con tales requisitos de transparencia y con la legislación de la UE en materia de derechos de autor, revelando que el contenido ha sido generado por IA y con la obligación de diseñar modelos que eviten la creación de contenidos ilegales.

La ley 2024/1689 quiere garantizar que la IA que vaya a utilizarse en el ámbito de la UE sea segura, transparente, trazable, no discriminatoria y respetuosa con el medioambiente.

4. Riesgo mínimo o nulo: en esta categoría, a la cual pertenecen la mayoría de los sistemas utilizados en la actualidad en la UE, se incluyen las aplicaciones como videojuegos habilitados para IA o los filtros de *spam.* La ley permite el uso de esta tecnología de riesgo mínimo.

Además, la ley impone varias obligaciones de transparencia y gestión de riesgos para los desarrolladores y proveedores de sistemas de IA. Por ejemplo, los sistemas que interactúan directamente con las personas deben estar claramente identificados como tales. También se requiere que los proveedores de modelos de esta tecnología de propósito general mantengan documentación técnica y resúmenes detallados sobre el contenido utilizado para el entrenamiento del modelo. Asimismo, la norma prevé la creación de mecanismos para fomentar la innovación, incluyendo espacios regulatorios de pruebas (*regulatory sandboxes*) donde se pueden desarrollar y probar sistemas de IA en condiciones controladas

El texto legislativo –que se compone de 180 considerandos, 113 artículos y 13 anexos– fue aprobado por el Parlamento Europeo en marzo del 2024, publicado en el *Diario Oficial de la Unión Europea* el 12 de julio del mismo año y entró en vigor el 1 de agosto. A partir de esa fecha, la entrada en vigor de las distintas disposiciones se producirá de forma gradual durante los siguientes 6 a 36 meses, y será de plena aplicación en todos los Estados miembros en un periodo de 24 meses desde su respectiva activación.

Finalmente, la Ley de IA establece nuevas estructuras de gobernanza, como una Oficina Europea de IA dentro de la CE y un Comité Europeo de Inteligencia Artificial para asegurar la aplicación coherente de las normas en toda la UE. La supervisión del cumplimiento y la aplicación de la ley quedará a cargo de las autoridades nacionales, apoyadas por la mencionada Oficina. Corresponderá a los Estados miembros la tarea de crear agencias nacionales de supervisión, para lo cual dispondrán un plazo de 12 meses después de su entrada en vigor el 2 de agosto de 2025 (Consejo de la Unión Europea 2024).

Certificaciones y estándares

Pese a que, en general, las empresas no participan directamente en el diseño y el desarrollo de los sistemas de IA, recae sobre ellas la responsabilidad de asegurarse de que la que usan sea ética y responsable. Para ello, tal como hemos visto, no solo pueden tomar como referencia la Ley de IA de la UE, sino toda la variedad de propuestas, recomendaciones y orientaciones que existen en este ámbito. Casi todas ellas mencionan principios como la privacidad, la equidad o no discriminación, la transparencia, la seguridad y la rendición de cuentas, por lo que parece haber consenso en que representan las principales características precisas para poder considerar tales sistemas como éticos.

No obstante, si bien los rasgos que son preferibles y deseables resultan fácilmente identificables, el desafío reside en las dificultades existentes para comprobar y tener garantías de que los sistemas de IA en cuestión cumplen, efectivamente, con esos principios. En este contexto, aparecen toda una serie de certificaciones y estándares de IA ética, que se articulan en torno a los marcos éticos de referencia mencionados. A través de ellos, se realizan evaluaciones rigurosas de los sistemas conforme a su grado de cumplimiento de los diferentes principios éticos en cuestión.

Así pues, mediante estas certificaciones y estándares de referencia que reconocen como éticos y responsables a los sistemas de IA, las empresas tienen a su alcance asegurarse de utilizar unos que cumplan con distintos principios a nivel ético. Con ese fin, a continuación se revisan algunos de los estándares y certificaciones más reconocidos que pueden ser considerados por las compañías a la hora de identificar aquellos que cumplen con altos niveles de ética y responsabilidad.

a) El Programa de Certificación Responsable de inteligencia artificial del Responsible AI Institute (RAII)

El Responsible AI Institute (RAII) es una organización sin ánimo de lucro que forma parte de la Alianza de Acción Global de IA (GAIA) del Foro Económico Mundial, dedicado a promover un uso responsable de esta tecnología. Desde el RAII se han desarrollado toda una serie de herramientas de análisis y estándares de referencia, y cuenta con de toda una red de expertos responsables en IA, con el fin de orientar tanto a las empresas como a los formuladores de políticas en la implementación de tecnologías que mejoren tanto el bienestar de las personas como el económico. Cuenta con diferentes tipos de evaluaciones, a través de las cuales las compañías desarrolladoras de sistemas de IA no solo pueden conocer los riesgos de sus productos, sino también demostrar a sus clientes el grado de responsabilidad que cumplen estos. Estas evaluaciones también son utilizadas con frecuencia por las compañías que planean implementar esos sistemas en sus actividades para asegurarse de que estos son responsables.

Aunque las empresas no participan en el diseño y el desarrollo de los sistemas de IA, tienen la responsabilidad de asegurarse de que los que usan sea éticos y responsables.

Sin embargo, en lo que destaca de forma especial la propuesta del RAII es en la creación de su propio programa de certificación independiente para una IA responsable. Dicha certificación, pese a ser de carácter independiente, se basa en diversos estándares, directrices y otros principios y políticas claves a nivel global, entre los que se otorga un peso especial a los "Principios de IA" de la OCDE, que incorporan objetivos de derechos humanos y buenas prácticas tecnológicas, y hacen un especial hincapié en la rendición de cuentas y la supervisión.

Ese proceso de certificación consiste, principalmente, en una auditoría realizada de forma independiente en la que se comprueba la alineación con toda una serie de requisitos. Dado que esta certificación reconoce la diversidad de significados de la IA, considera que no puede establecerse un único programa de certificación válido para todos los sistemas. Además, incorpora consideraciones relativas al contexto del caso de uso, la industria y la región del sistema de IA.

Por el momento, la certificación RAII únicamente es aplicable en los siguientes casos de uso, si bien la intención es aumentar esta lista:

- Préstamos automatizados (finanzas)
- Cobros automatizados (finanzas)
- Adquisiciones (todas las industrias)
- Recursos humanos (todas las industrias)
- Acceso a la atención sanitaria (cuidado de la salud)
- Imágenes de la piel (cuidado de la salud)

La evaluación que se lleva a cabo en este programa de certificación consiste, en concreto, en un conjunto de 89 preguntas, indicadores de respuesta y requisitos de evidencia con los que se trata de medir la madurez de la IA a nivel del sistema, abarcando las siguientes 6 dimensiones, con 20 subdimensiones:

- Operaciones de sistemas: se explora el funcionamiento del sistema de IA y las opciones de diseño clave relacionadas con el modelo y sus datos.
- Explicabilidad e interpretabilidad: se comprueba que el funcionamiento y los usos del sistema de IA puedan explicarse y documentarse en términos que las personas puedan entender.
- Responsabilidad: se examina si la organización ha establecido procesos de supervisión claros para el desarrollo y la implementación del sistema de IA.
- Protección al consumidor: se evalúa el riesgo que el sistema de IA representa para las personas y los pasos que la organización y el equipo de desarrollo han dado para mitigarlo.
- Sesgo y equidad: se examina si el sistema de IA fue diseñado de una manera que promueva la equidad y evite sesgos.
- Robustez: investiga si el sistema de IA es seguro y eficaz.

A su vez, para las respuestas a cada pregunta se proporcionan distintas puntuaciones, cuyo cómputo global determina el nivel de certificación que se puede atribuir al sistema de IA evaluado. De nuevo, cabe insistir en que, a lo largo de estas respuestas, desde el RAII se tienen en consideración el modelo y la implementación contextual de un sistema de IA, así como la interacción del dominio de este, la región y el tipo concreto de que se trata. A partir de ello, las respuestas a las preguntas de evaluación se califican según este sistema de puntuación:

0: Necesita mejorar
1: Satisfactorio
3: Bueno
5: Excelente

Si un sistema de IA obtiene más del 50% de la puntuación disponible en cada una de las dimensiones, se suma la puntuación de estas para obtener la total de la evaluación. Entonces, esta última se representa en forma de porcentaje, a partir del cual se determina el nivel de certificación del sistema de IA. En la **Figura 5** pueden observarse los distintos porcentajes de puntuación de la evaluación y sus correspondientes niveles de certificación.

Figura 5. Niveles de certificación del RAII

Total Score	Level Obtained	Corresponding Mark
0-49.9%	Not Certified	N/A
50-59.9%	Certified	
60-69.9%	Silver	SILVER
70-79.9%	Gold	GOLD
80+%	Platinum	PLATINUM

Fuente: RAII (2022).

b) La Organización Internacional de Estandarización (ISO) y la Comisión Electrotécnica Internacional (IEC)

La Organización Internacional de Estandarización (ISO) está realizando de forma conjunta con la Comisión Electrotécnica Internacional (IEC) grandes esfuerzos de estandarización en el ámbito de la IA. Para ello, en el 2017 desarrollaron el Comité de Estándares Internacionales conocido como *ISO/IEC JTC 1/ SC42*, desde el que se analiza de forma específica todo el ecosistema de la IA y se proporciona orientación y coordinación para los subcomités ISO e IEC que desarrollan estándares para ella. Desde esta instancia se han generado toda una serie de estándares que abarcan distintas dimensiones de la transparencia, la calidad de los datos y la confiabilidad de los sistemas de IA.

Así pues, el cumplimiento de dichos estándares es indicativo de contar con altos grados de responsabilidad y buen hacer, por lo que pueden resultar un medio a través del cual las empresas identifiquen sistemas de IA éticos y responsables.

Las certificaciones brindan a las empresas la capacidad de identificar y seleccionar sistemas de IA éticos y responsables.

Entre los estándares desarrollados por el ISO/IEC JTC 1/ SC42 destacan los siguientes:

- ISO/IEC 22989:2022. Enfocado en establecer una terminología para la IA estandarizada, proporciona conceptos unificados con el fin de ayudar a que un conjunto más amplio de partes interesadas comprenda y utilice mejor esta tecnología. En este sentido, se destina a un público amplio, en el que se incluyen tanto expertos como no profesionales.
- ISO/IEC 23053:2022. Centrado particularmente en el ML, rama de la IA que, a través de técnicas computacionales, permite que los sistemas aprendan de datos o experiencias, de manera que puedan realizar tareas específicas de forma autónoma, sin necesidad de ser programados. En este sentido, trata de proporcionar una descripción conjunta del ML, estableciendo una terminología común para dichos sistemas, lo cual proporciona una base conjunta que permite la explicación clara de estos y diversas consideraciones que se aplican a su ingeniería y su uso, así como para otros estándares dirigidos a aspectos específicos de los sistemas de ML y sus componentes.
- ISO/IEC 42001:2023. Tiene como objetivo ayudar a las organizaciones a desempeñar de un modo responsable su función con respecto a los sistemas de IA, tanto si es haciendo uso de ellos como desarrollándolos, monitoreándolos o proporcionando bienes o servicios que los utilicen. Ofrece orientación para establecer, implementar, mantener y mejorar continuamente un sistema de gestión de IA dentro del contexto de una organización. Por tanto, es aplicable por parte de cualquier organización, con independencia de su tamaño, tipología o sector al que pertenezca, que utilice o suministre productos que empleen sistemas de IA.

Los objetivos de la norma ISO/IEC 42001:2023 se pueden sintetizar de la siguiente manera:

- Fomentar el desarrollo y la implementación de sistemas de IA confiables, transparentes y responsables.
- Asistir a las organizaciones en la identificación y mitigación de los riesgos asociados con la adopción de la IA, lo que a su vez mejora la eficacia y reduce los costes.
- Fomentar una mayor confianza en la gestión de la IA al alentar a las organizaciones a priorizar el bienestar humano, la seguridad y la experiencia del usuario en el diseño e implementación de esta tecnología.

Asimismo, este estándar tiene como objetivo ayudar a la organización a desarrollar, proporcionar o utilizar sistemas de IA de manera responsable, con el fin de lograr un equilibrio entre alcanzar sus objetivos y cumplir con los requisitos reglamentarios y las obligaciones relacionadas con las partes interesadas y las expectativas de estas.

- ISO/IEC 23894:2023. Proporciona orientación sobre cómo las empresas que desarrollan, producen, implementan o utilizan bienes, sistemas y servicios que utilizan IA pueden gestionar los riesgos que derivan propiamente de esta (riesgos inherentes). Por tanto, trata de orientar a las organizaciones en la integración de la gestión de riesgos en sus actividades y funciones relacionadas con la IA, mediante la descripción los procesos necesarios para ello.

La aplicación de esta guía es personalizable para cualquier organización y su contexto, y actúa como complemento de la ISO 31000:2018, enfocada en proporcionar directrices para gestionar de forma general los riesgos a los que se enfrentan las organizaciones, ya que la ISO/IEC 23894:2023 se centra en las consideraciones específicas que requiere la IA con respecto a los principios descritos en la ISO 31000:2018.

- ISO/IEC TR 24027:2021. Enfocado en abordar los riesgos por posibles sesgos en los sistemas de IA, brinda a las organizaciones las orientaciones necesarias con las que poder identificarlos y tratarlos, garantizando así que, en última instancia, las partes interesadas puedan beneficiarse de los sistemas de IA de acuerdo con sus objetivos.

Para ello, por un lado, ofrece una visión general de los posibles sesgos no deseados, así como de sus fuentes potenciales. Por otro lado, proporciona orientaciones, técnicas y métodos de medición con los que poder evaluar los grados de sesgo y equidad. Con ese fin, se proponen diferentes sistemas de medición para las distintas fases del ciclo de vida del sistema de IA: des-

de la recopilación de datos hasta la capacitación, pasando por el aprendizaje continuo, el diseño, las pruebas, la evaluación y el uso. Por último, se ofrece un catálogo de distintas posibles estrategias de tratamiento de los sesgos identificados.

- ISO/IEC 27001:2022. Proporciona a las organizaciones un marco integral para gestionar los riesgos asociados con la seguridad de datos. Con ese objetivo, establece una serie de requisitos para establecer, implementar, mantener y mejorar continuamente los sistemas de gestión de la seguridad de la información en el contexto de la organización, a cuyos efectos incluye requisitos para la evaluación y el tratamiento de los riesgos relativos a estas cuestiones.

El cumplimiento de la ISO/IEC 27001:2022 representa el compromiso de una organización de adoptar las mejores prácticas y estándares internacionales para salvaguardar activos de información de naturaleza crítica. Por tanto, puede ser utilizada por partes tanto internas como externas a la hora de evaluar la capacidad de un sistema de IA para cumplir con ciertos requisitos de seguridad de la información:

- ISO 31700-1:2023. Se basa en la idea de privacidad por diseño, la cual hace referencia a la idea de que los usuarios no tengan que soportar la carga de luchar por la protección de su privacidad cuando utilizan productos de consumo. Así pues, trata de incorporar este principio en el desarrollo de bienes, procesos, sistemas, *software* y servicios, con el fin de que se tenga en cuenta la privacidad de un consumidor durante todo el ciclo de vida de un producto. Es decir, significa que un producto tiene por defecto controles y configuraciones de privacidad predeterminados orientados al consumidor que brindan niveles apropiados de privacidad, sin imponer una carga indebida a este.
- Al respecto, la ISO 31700-1:2023 establece los requisitos y procesos necesarios a través de los cuales poder incorporar y asegurar esta privacidad por diseño en aquellos productos en los que haya presencia de IA, por lo que se dirige especialmente al personal de las organizaciones y terceros responsables del concepto, diseño, fabricación, gestión, pruebas, operación, servicio, mantenimiento y eliminación de bienes y servicios de consumo.

c) El Institute of Electrical and Electronics Engineers (IEEE)

El IEEE es una asociación internacional sin ánimo de lucro conformada por ingenieros eléctricos y electrónicos dedicada al avance de la tecnología en beneficio de la humanidad. Se trata de una de las organizaciones líderes del mundo en la creación de estándares. De hecho, en la actualidad cuenta con más de 1.900, que se enfocan en una amplia gama de industrias. En el contexto del tema al que se dedica este cuaderno, el IEEE puso en marcha en el año 2019 la Iniciativa Global para la Ética de los Sistemas Autónomos e Inteligentes, con el fin de abordar de forma específica los problemas éticos relacionados con la creación y la difusión de los sistemas de IA. Esta iniciativa persigue, básicamente, la educación y la capacitación de todas las partes interesadas involucradas en el diseño y el desarrollo de sistemas autónomos e inteligentes, con el fin de que incorporen consideraciones éticas en sus acciones.

Dentro de este proyecto, el IEEE ha desarrollado todo un conjunto de estándares conocidos como ***serie*** *IEEE P7000™*, que se enfocan, de forma específica, en el ámbito de los sistemas de IA. En la actualidad, se han desarrollado 15 estándares dentro de este marco, que se muestran en la **Tabla 1:**

Tabla 1. Familia de estándares P7000™ desarrollados por el IEEE y ámbito que abordan

IEEE P7000™	Preocupaciones éticas durante el diseño del sistema
IEEE P7001™	Transparencia de sistemas autónomos
IEEE P7002™	Proceso de privacidad de datos
IEEE P7003™	Consideraciones de sesgo algorítmico
IEEE P7004™	Norma sobre gobernanza de datos de niños y estudiantes
IEEE P7005™	Norma sobre gobernanza de datos de los empleadores
IEEE P7006™	Grupo de trabajo de agentes de IA de datos personales
IEEE P7007™	Norma ontológica para sistemas de automatización y robótica impulsados éticamente
IEEE P7008™	Estándar para impulsar éticamente sistemas robóticos, inteligentes y autónomos
IEEE P7009™	Norma para el diseño a prueba de fallos de sistemas autónomos y semiautónomos
IEEE P7010™	Métricas de bienestar para sistemas autónomos e inteligentes
IEEE P7011™	Proceso de identificación y calificación de la confiabilidad de fuentes de noticias
IEEE P7012™	Términos de privacidad personal legibles por máquina
IEEE P7013™	Estándares de inclusión y aplicación en tecnología de análisis facial automatizado
IEEE P7014™	Estándar para consideraciones éticas en empatía emulada en sistemas autónomos e inteligentes

De este modo, cualquier sistema de IA que se encuentre vinculado o que acredite cumplir con cualquiera de estos estándares es indicativo de que cuenta con un alto nivel de compromiso y de ética en algún aspecto concreto.

Los estándares ISO evalúan distintas dimensiones de la transparencia, la calidad de los datos y la confiabilidad de los sistemas de IA.

d) El AI Ethics Impact Group (AIEIG)

El AIEIG es un consorcio interdisciplinar liderado por la VDE Association for Electrical, Electronic & Information Technologies y Bertelsmann Stiftung. A la luz de la gran acogida y expansión que los sistemas de IA han experimentado, este grupo ha optado por desarrollar un marco propio con el que pretende mostrar cómo poner en práctica principios éticos en el campo de esta tecnología y cómo evaluar los distintos sistemas de acuerdo con esos principios, a través de un modelo conocido como *VCIO*, que distingue y combina cuatro conceptos: valores, criterios, indicadores y observables (de ahí sus siglas).

Asimismo, propone la existencia de una etiqueta de ética de la IA, inspirada en la de eficiencia energética, con la que se aumentaría la transparencia y la comparabilidad de los productos para los usuarios y se proporcionaría una base para una mejor supervisión por parte de los formuladores de políticas, los reguladores, las asociaciones de desarrollo de estándares y las organizaciones de vigilancia.

La evaluación de los sistemas de IA mediante el modelo VCIO se basa en estos en seis aspectos:

- **Transparencia:** en esta categoría se considera tanto la explicabilidad técnica del algoritmo como la transparencia del proceso de desarrollo y capacitación del sistema. La explicabilidad consiste en que los usuarios puedan comprender completamente el "funcionamiento interno", lo cual permite resolver problemas de atribución, es decir, la capacidad de demostrar por qué se han producido determinados errores. Por su parte, la transparencia en el proceso de desarrollo se refiere a cuestiones concretas: quién es responsable del desarrollo de los modelos de ML, quién aprobó este, de dónde provienen los datos utilizados para entrenar los modelos, a qué pruebas de calidad se han sometido los conjuntos de datos y quién los ha etiquetado, qué

objetivos de aprendizaje se persiguen, qué resultados se obtienen con las evaluaciones de los modelos, qué métodos de aprendizaje se utilizan, etc.

- **Rendición de cuentas:** en esta categoría se evalúa el modo en que un sistema de IA asume la voluntad y obligación de responsabilizarse. Es decir, por un lado, se comprueba en qué medida el sistema especifica sus obligaciones y, por otro, si tiene designado a un agente responsable. De esta forma se contrarresta la posible difusión y fuga de responsabilidades. En este sentido, deben designarse personas o instituciones que serán las responsables de cumplir con las obligaciones pertinentes e identificables por todos los afectados, así como fácilmente accesibles para las posibles preguntas, quejas o apelaciones. Asimismo, es fundamental que instituciones, empresas o individuos se responsabilicen financieramente de los sistemas de IA y de las formas de reparación no monetarias correspondientes.

La iniciativa de la IEEE persigue la correcta capacitación de las personas involucradas en el diseño y el desarrollo de sistemas de IA.

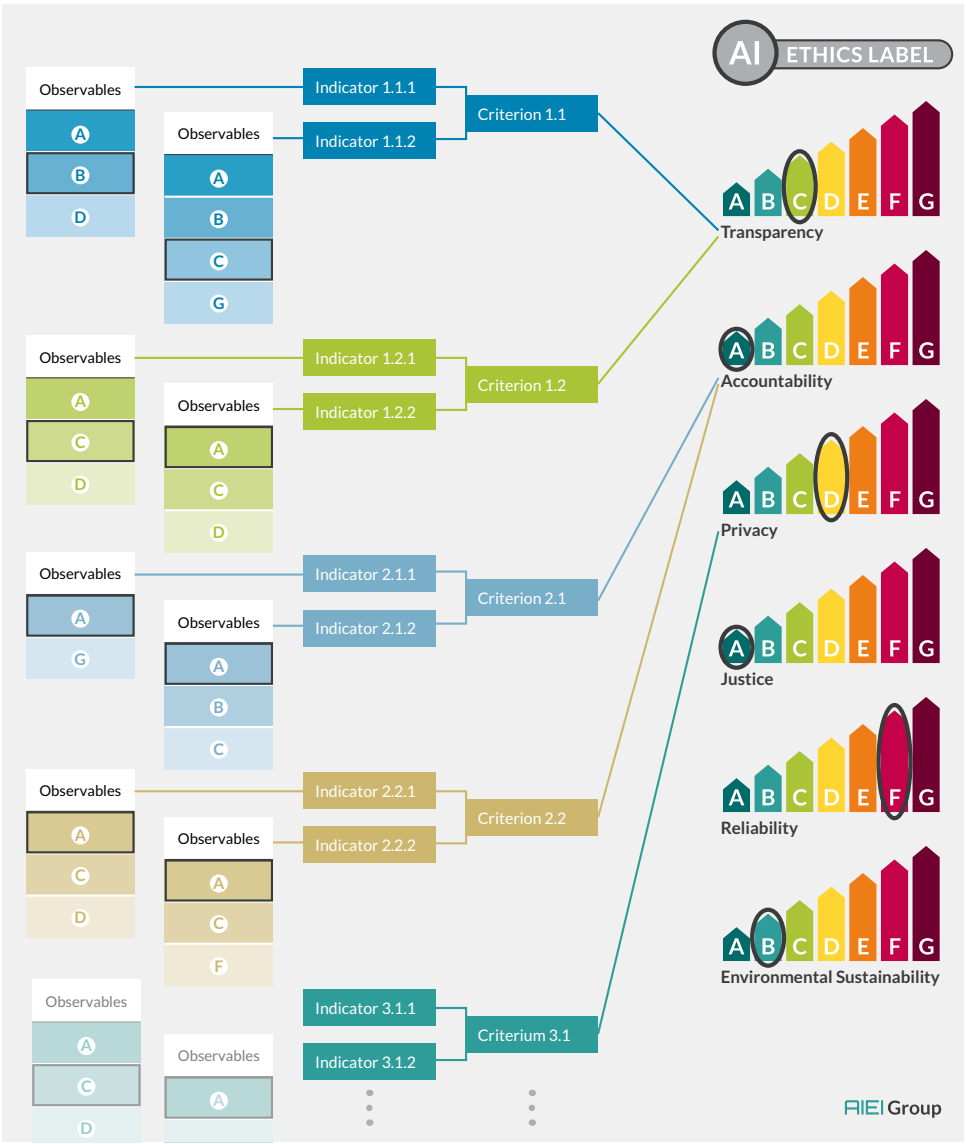
- **Privacidad:** según el AIEIG, para garantizar la privacidad, la ética de la IA debe considerar varios principios básicos sobre el uso y la gestión de los datos: los datos personales solo pueden recopilarse y utilizarse para fines específicos; una vez cumplida la finalidad, no podrán seguir siendo tratados, y solo podrán emplearse para una finalidad distinta de aquella para la que fueron recogidos, si los interesados han dado su consentimiento explícito. La privacidad también incluye el principio de consentimiento, es decir, el derecho a que sean eliminados o rectificados, y la capacidad de restringir su procesamiento, lo que significa que las personas tienen el poder de impedir que sus datos sean utilizados en aplicaciones de IA. Además, en el desarrollo de sistemas de IA debe promoverse el aprendizaje con datos anónimos o seudónimos y el uso de procedimientos fiables de anonimización y seudonimización, por lo que en esta categoría también se examina el potencial de anonimización de tales sistemas.
- **Justicia:** en esta categoría se evalúa si existen problemas de igualdad de trato o discriminación algorítmica. Se refiere, por un lado, a distinciones y clasificaciones que no tendrían que desempeñar un papel en la acción porque se basan en estereotipos o atribuciones degradantes, o en atributos que no deberían tener un impacto material en una decisión. Es el caso de categorías como el género, la edad, el origen étnico o la nacionalidad, la discapacidad o el embarazo. Por otro lado, hace referencia a la posible reproducción de patrones de discriminación existentes que se introducen a través de los datos de entrenamiento, dando lugar a un trato diferente e injustificado de determinados grupos a nivel de aplicación. Asimismo, como novedad, este modelo de evaluación también contempla aspectos de justicia social relacionados con el trabajo "oculto" que implica el funcionamiento de los sistemas de IA. Es decir, comprueba que el desarrollo de estos no incluya mano de obra precaria, actividades perjudiciales para la salud de sus trabajadores o cualquier incumplimiento de derechos laborales.
- **Fiabilidad:** en la medida en que las consecuencias de resultados erróneos, accidentes o mal uso de los sistemas de IA pueden afectar a individuos, partes de un sistema o a toda una sociedad, es necesario desarrollar estrategias adecuadas para construir una infraestructura confiable. En este sentido, las aplicaciones de esta tecnología se consideran confiables cuando funcionan de la manera prevista y cuando no poseen vulnerabilidades frente a atacantes externos, de modo que pueden evitar manipulaciones de diversos tipos. Asimismo, también se refiere a la precisión y reproducibilidad de los resultados del sistema y a la capacidad de resistir interferencias.
- **Sostenibilidad ambiental:** en esta categoría se evalúa el grado en que un sistema de IA favorece un uso cuidadoso de los recursos naturales, por ejemplo, en las infraestructuras o los consumos de energía que requieren. Además de los impactos negativos significativos, también se miden y valoran los efectos positivos en el medioambiente. A pesar de que la IA brinda grandes soluciones sostenibles a muchos de los desafíos climáticos actuales, se trata de una tecnología cuyo desarrollo e implementación también implican una serie de impactos ambientales que deben ser considerados. Es el caso, por ejemplo, de los centros de datos a través de los cuales operan los sistemas de IA, infraestructuras que suponen un gran consumo energético. En concreto, en el 2022 representó en la UE, según algunas estimaciones, un 4% del consumo total (AIE 2024). Asimismo, estos centros de datos son también responsables de una gran cantidad de emisiones de GEI, ya que la mayoría de ellos se alimentan de fuentes de energía que todavía dependen mayoritariamente de combustibles fósiles. Según la AIE (s. f.), los centros de datos y las redes de transmisión de estos son responsables del 1% de dichas emisiones. Ante el potencial crecimiento del sector y de sus impactos, desde el marco del Pacto Verde Europeo se ha establecido especí-

ficamente el objetivo de que, para el 2025, el 70% de la energía que adquieran los centros de datos proceda de fuentes renovables, con vistas a alcanzar el 100% en el 2030 (Climate Neutral Data Center, s. f.).

A partir de la evaluación de estos valores mediante el método VCIO, los resultados quedarían traducidos de un modo tal que los ciudadanos, usuarios y consumidores podrían entender fácilmente. Además, la etiqueta de ética propuesta incluye una calificación para cada uno de los valores capturados conforme a su nivel de cumplimiento en un sistema y un gráfico de puntuación fácilmente reconocible, tal como se refleja en la **Figura 6.**

El desarrollo e implementación de la IA también implican una serie de impactos ambientales que deben ser considerados.

Figura 6. Proceso de calificación de la etiqueta de ética de la IA



Fuente: Hallensleben y Hustedt (2020).

Conclusión

La IA promete grandes avances que, tal como se ha visto en este cuaderno, si se canalizan hacia el bien de las personas tiene el potencial de mejorar la calidad de nuestras vidas de un modo sin precedentes. No obstante, el desarrollo de los sistemas de IA presenta toda una serie de desafíos en cuanto a su diseño y uso, que suscitan algunas dudas tanto entre los distintos agentes implicados como entre la opinión pública.

En cuanto a su implementación, si bien esta es amplia en cuestiones relativas a la mejora de la eficiencia de tareas o en términos de entretenimiento, en el ámbito económico y laboral es considerablemente inferior debido a las no pocas preocupaciones que despierta, al igual que lo hace en el aspecto de la privacidad de los datos. En respuesta a estas inquietudes, se han promovido diferentes regulaciones que persiguen garantizar un uso correcto de esta tecnología. Al respecto, la reciente Ley de IA la UE (2024) ha marcado un hito importante en esta dirección.

En el caso concreto del sector empresarial, la IA también está ganando terreno debido a los grandes beneficios que ya están experimentando las organizaciones que la están implementando para la realización de actividades diversas. Sin embargo, el riesgo de comisión de errores que puedan perjudicar gravemente su reputación es, entre otros, un motivo de peso por el que muchas de ellas se muestran reticentes a incorporar esta tecnología. En este sentido, al tratarse de una materia en continuo desarrollo, se hace difícil definir la forma concreta en que aquellas deben implementar estas ideas y la manera en que pueden garantizar un uso adecuado de la IA. Por ello, acudir a certificaciones y estándares elaborados, en su gran mayoría, por grupos de expertos, en consonancia con las principales directrices éticas representa, hoy en día, un medio más que fiable a través del cual las compañías tienen a su alcance asegurarse de que la implantación de sistemas de IA es acorde a toda una serie de principios éticos a partir de procedimientos claros y definidos.

En cualquier caso, lo que resulta innegable –e invita, en cierta manera, al optimismo– es que las potenciales bondades económicas que brinda esta tecnología no están dejando de lado la preocupación por el bienestar de la sociedad. En este sentido, parece que estamos optando por asegurar un desarrollo tecnológico a partir de unos requisitos mínimos con el objetivo de no comprometer nuestras formas de vida ni vulnerar nuestros derechos más fundamentales. El hecho de no precipitarse hacia lo desconocido, lo cual podría provocar toda una serie de consecuencias negativas en el bienestar de las personas, y la priorización de un desarrollo precavido y responsable son, sin duda, grandes noticias. Sin embargo, la responsabilidad de lograr que esta tendencia se mantenga recae, en gran parte, sobre las empresas, que deben seguir explorando estas novedades tecnológicas bajo esta línea de desarrollo consciente y ética.

Por último, si bien la aparición de un marco normativo en este ámbito se configura como un hito fundamental –y así lo hemos reflejado en este cuaderno–, cabe insistir en que la regulación no debe ser percibida como el único medio para asegurar un desarrollo y un uso ético de la IA. La palanca de las normas y de la regulación –y las certificaciones y los estándares que de ellas puedan derivarse– ha de apoyarse y complementarse con otras medidas, en especial con las que definan unos valores y unas competencias morales en los decisores (tanto quienes intervienen en el diseño e implementación de los sistemas de IA como los propios usuarios) para asegurar un uso ético de esta tecnología. De todas estas cuestiones hablaremos en futuros cuadernos.

Referencias

AIE (Agencia Internacional de la Energía). 2024. *Electricity 2024. Analysis and Forecast to 2026.* https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf.

AIE (Agencia Internacional de la Energía). s. f. "Data Centres and Data Transmission Networks". Acceso el 30 de octubre del 2024. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.

BBC Mundo. 2018. "5 claves para entender el escándalo de Cambridge Analytica que hizo que Facebook perdiera US$37.000 millones en un día". Actualizado el 21 de marzo. https://www.bbc.com/mundo/noticias-43472797.

CLIMATE Neutral Data Center. s. f. "Climate Neutral Data Centre Pact". Acceso el 30 de octubre del 2024. https://www.climateneutraldatacentre.net/#twae-scrollar-dd105e6-item-5.

COMISIÓN Europea. 2019. *Directrices éticas para una IA fiable.* Dirección General de Redes de Comunicación, Contenido y Tecnologías. https://data.europa.eu/doi/10.2759/14078.

CONSEJO de la Unión Europea. 2024. "Artificial Intelligence (AI) Act: Council Gives Final Green Light to the First Worldwide Rules on AI". Comunicado de prensa. 21 mayo. https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/.

DASTIN, Jeffrey. 2018. "Insight- Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women". Reuters. 11 de octubre. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/.

ECONOMIC Times. 2023. "Uber's Surge Pricing Sparks Controversy amidst Record-breaking Profits". Actualizado el 2 de agosto. https://economictimes.indiatimes.com/news/new-updates/ubers-surge-pricing-sparks-controversy-amidst-record-breaking-profits/articleshow/102344986.cms?from=mdr.

EY. 2024. *How Can You Realize the Promise of Transformational Technologies? EY Reimagining Industry Futures Study 2024.* https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/tmt/ey-reimagining-industry-futures-study-2024-report.pdf.

FORO Económico Mundial. 2023. *Data Equity: Foundational Concepts for Generative AI.* https://www3.weforum.org/docs/WEF_Data_Equity_Concepts_Generative_AI_2023.pdf.

HAAN, Katherine y Rob Watts. 2023. "How Businesses Are Using Artificial Intelligence in 2024". *Forbes Advisor.* Actualizado el 24 de abril. https://www.forbes.com/advisor/business/software/ai-in-business/.

HALLENSLEBEN, Sebastian y Carla Hustedt. 2020. *From Principles to Practice: An Interdisciplinary Framework to Operationalise AI Ethics. AI Ethics Impact Group.* https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf.

IBM Consulting. 2023. "Artificial Intelligence and a New Era of Human Resources". IBM. 9 de octubre. https://www.ibm.com/blog/artificial-intelligence-and-a-new-era-of-human-resources/.

IBM Data and AI Team. 2023. "Understanding the Different Types of Artificial Intelligence". IBM. 12 de octubre. https://www.ibm.com/think/topics/artificial-intelligence-types.

IBM Institute for Business Value. 2024. *Revolutionize retail with AI everywhere. IBM.* https://www.ibm.com/downloads/cas/35BVNBNA.

IPSOS. 2023. *Global views on AI 2023. How People across the World Feel about Artificial Intelligence and Expect It Will Impact Their Life.* https://www.ipsos.com/sites/default/files/ct/news/documents/2023-07/Ipsos%20Global%20AI%202023%20Report-WEB_0.pdf.

KENNEDY, Brian, Alec Tyson y Emily Saks. 2023. "Public Awareness of Artificial Intelligence in Everyday Activities". Pew Research Center. 15 de febrero. https://www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities/.

MASLEJ, Nestor, Loredana Fattorini, Raymond Perrault *et al. 2024. The AI Index 2024 Annual Report. Institute for Human-Centered AI, Universidad de Stanford.* https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf.

MCCARTHY, J., M. L. Minsky, N. Rochester y C. E. Shannon. 1955. *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Formal Reasoning Group, Universidad de Stanford.* https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

MICROSOFT Azure. s. f. "¿Qué es la inteligencia artificial?". Microsoft. Acceso el 30 de octubre del 2024. https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-is-artificial-intelligence#autom%C3%B3viles-sin-conducto.

MORANDÍN-AHUERMA, Fabio. 2023. *Principios normativos para una ética de la Inteligencia Artificial.* Consejo de Ciencia y Tecnología del Estado de Puebla (México). https://philpapers.org/archive/MORIUE-2.pdf.

NAVEEN, Joshi. 2019. "7 Types of Artificial Intelligence". *Forbes.* 19 de junio. https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/?sh=5629badd233e.

OCDE. 2024. *Recommendation of the Council on Artificial Intelligence.* OECD/LEGAL/0449. 22 de mayo. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449.

PAZZANESE, Christina. 2020. "Great promise but potential for peril". *Harvard Gazette*. 26 de octubre. https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/.

QUACH, Katyanna. 2023. "Corporate Investment in AI down for First Time in a Decade". *The Register*. 5 de abril. https://www.theregister.com/2023/04/05/corporate_investment_in_ai_drops/.

QUANTUMBLACK, AI by McKinsey. 2024. *The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value.* https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2024/the-state-of-ai-in-early-2024-final.pdf?shouldIndex=false.

RAII (Responsible Artificial Intelligence Instu¡itute). 2022. *The Responsible AI Certification Program.* https://20965052.fs1.hubspotusercontent-na1.net/hubfs/20965052/RAII%20Certification%20White%20Paper.pdf.

REGLAMENTO (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.° 300/2008, (UE) n.° 167/2013, (UE) n.° 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial). 2024. *Diario Oficial de la Unión Europea. 12 de julio*. https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=OJ:L_202401689.

SHERMAN, Len. 2023. "Uber's New Math: Increase Prices and Squeeze Driver Pay". *Forbes.* 16 de enero. https://www.forbes.com/sites/lensherman/2023/01/16/ubers-new-math-increase-prices-and-squeeze-driver-pay/?sh=1ba22e24c8a2.

SHRM (Society for Human Resource Management). 2024. *2024 Talent Trends: Artificial Intelligence in HR*. https://shrm-res.cloudinary.com/image/upload/AI/2024-Talent-Trends-Survey_Artificial-Intelligence-Findings.pdf.

TURING, Alan. 1950. "Computing Machinery and Intelligence". *Mind.*

UBER. s. f. "How surge pricing works". Acceso el 30 de octubre del 2024. https://www.uber.com/us/en/drive/driver-app/how-surge-works/.

UNESCO. 2021. *Recomendación sobre la ética de la inteligencia artificial.* 23 de noviembre. https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa.

www.iese.edu

Barcelona
Madrid
Munich
New York
São Paulo

IESE Business School
University of Navarra

A Way to Learn. A Mark to Make. A World to Change.