Universidad de Valladolid

ESCUELA DE INGENIERÍAS INDUSTRIALES

UNIVERSIDAD DE VALLADOLID

ESCUELA DE INGENIERIAS INDUSTRIALES

Grado en Ingeniería Electrónica Industrial y Automática

El papel de la inteligencia artificial en la sociedad: retos y posibilidades

Autor:

Sanz Casado, María

Tutora:

Quintano Pastor, María del Carmen

Departamento Tecnología Electrónica

Valladolid, junio 2025

Resumen:

Este Trabajo de Fin de Grado analiza el papel de la inteligencia artificial en la sociedad actual desde una perspectiva de ingeniería. Se estudian las posibilidades que ofrece la IA en campos como la automatización de procesos, la toma de decisiones, la medicina, la educación y el desarrollo industrial, así como su impacto en la eficiencia y la innovación tecnológica. A su vez, se identifican los principales riesgos y desafíos éticos que plantea su implementación, incluyendo cuestiones relacionadas con la privacidad, el empleo, la seguridad y la equidad social. El trabajo pone de relieve la necesidad de establecer marcos normativos claros que garanticen un desarrollo responsable y sostenible de esta tecnología. Finalmente, se reflexiona sobre las repercusiones sociales, económicas y ambientales de la IA, proponiendo un enfoque crítico y ético que permita maximizar sus beneficios minimizando sus efectos negativos.

Palabras clave: Inteligencia artificial, riesgos, sociedad, desafíos éticos, oportunidades.

Abstract:

This Final Degree Project analyzes the role of artificial intelligence today from an engineering perspective. It examines the opportunities AI offers in fields such as process automation, decision-making, healthcare, education, and industrial development, as well as its impact on efficiency and technological innovation. At the same time, it identifies the main risks and ethical challenges associated with its implementation, including issues related to privacy, employment, security, and social equity. The project highlights the need to establish clear regulatory frameworks to ensure responsible and sustainable development of this technology. Finally, it reflects on the social, economic, and environmental repercussions of AI, proposing a critical and ethical approach that maximizes its benefits while minimizing its negative effects.

Keywords: Artificial intelligence, risks, society, ethical challenges, opportunities.

Índice

1. Introducción

La inteligencia artificial (IA) está transformando progresivamente la manera en que interactuamos con la tecnología, accedemos a los servicios y organizamos los procesos económicos y sociales. Aunque su desarrollo comenzó hace varias décadas, en los últimos años se ha producido una aceleración sin precedentes, impulsada por la disponibilidad de grandes cantidades de datos, el aumento de capacidad de procesamiento y los avances en algoritmos de aprendizaje automático.

Hoy en día, la IA no solo está presente en sectores altamente tecnológicos, sino que también se ha integrado en actividades cotidianas como el consumo digital, la movilidad, la asistencia sanitaria o la educación. Esta expansión ha abierto múltiples posibilidades como la automatización de tareas, la optimización de recursos y la mejora en la toma de decisiones. Al mismo tiempo han surgido desafíos importantes relacionados con la ética, la responsabilidad, la equidad, el empleo, la privacidad y la gestión de estas tecnologías.

El debate académico sobre la IA se divide en dos grandes ejes. Por un lado, la literatura técnica que se enfoca en el desarrollo y mejora de modelos predictivos, algoritmos de aprendizaje, visión artificial, procesamiento del lenguaje natural y robótica autónoma. Por otro lado, una creciente línea de investigación interdisciplinar que aborda el impacto social, económico, ético y jurídico de esta tecnología.

Comprender el alcance y la complejidad del impacto de la inteligencia artificial en la sociedad se ha vuelto fundamental para evaluar sus efectos a corto y largo plazo. Este trabajo busca ofrecer un análisis crítico y estructurado sobre este fenómeno atendiendo tanto a sus bases teóricas, como a sus aplicaciones y consecuencias. El propósito es proporcionar una visión amplia, pero rigurosa que permita reflexionar sobre el papel que esta tecnología desempeña y seguirá desempeñando en nuestras vidas.

1.1 Objetivos del proyecto

El objetivo general de este trabajo es comprender y analizar la inteligencia artificial, desde sus bases teóricas hasta sus potenciales beneficios, considerando sus capacidades, limitaciones, aplicaciones y efectos sociales.

Los objetivos específicos son:

- Definir y explicar los principios esenciales de la inteligencia artificial.
- Evaluar el impacto social y psicológico de la inteligencia artificial en la sociedad.

- Analizar los riesgos éticos, psicológicos, laborales y medioambientales asociados al desarrollo de la inteligencia artificial.
- Mostrar el potencial transformador que la inteligencia artificial ofrece a la sociedad.

1.2 Motivación

La elección de este Trabajo de Fin de Grado surgió de un profundo interés sobre la revolución tecnológica que estamos viviendo. Es fascinante observar cómo la IA ha dejado de ser algo de ciencia ficción para convertirse en una presente realidad, redefiniendo la forma que nos relacionamos, trabajamos y hasta entendemos el mundo. Me interesa mucho entender cómo funciona por dentro, sí, pero sobre todo qué consecuencias tiene esto para nosotros.

Me llama especialmente la atención ver cómo la inteligencia artificial ya se aplica en campos tan importantes como la salud y la educación, mejorando procesos y abriendo nuevas puertas. Sin embargo, al mismo tiempo es evidente que su avance plantea dilemas serios: ¿Qué pasa con nuestros trabajos?, ¿Cómo afecta a nuestra privacidad?, ¿Qué impacto psicológico tiene?, o incluso ¿Qué coste tiene para el medioambiente? Abordar estas preguntas, buscando un equilibrio entre el enorme potencial de la IA y los desafíos qué tiene es lo que me motivó a investigar y a documentarme sobre este tema. Este trabajo representa una oportunidad para analizar críticamente tanto sus promesas como sus sombras, buscando ofrecer una visión equilibrada y así contribuir a un debate informado sobre cómo podemos guiar el desarrollo de la IA hacia un futuro que sea realmente beneficioso para todos.

1.3 Estructura del proyecto

Para abordar de manera clara y exhaustiva el tema del papel de la inteligencia artificial en la sociedad, este Trabajo de Fin de Grado se ha estructurado en seis capítulos interconectados. Tras una introducción donde se establece el contexto, la justificación y los objetivos de estudio, el segundo capítulo inteligencia artificial bases teóricas sienta los fundamentos detallando su historia, definiciones, tipos y principios operativos, con especial atención al Machine learning y al Deep learning. El impacto social de la inteligencia artificial se aborda en el tercer capítulo explorando su influencia en diversos sectores como el empleo, la educación o el entretenimiento. El cuarto capítulo retos, peligros y desafíos de la Inteligencia Artificial analiza en profundidad los riesgos actuales y potenciales de la IA, incluyendo aspectos éticos, de seguridad, medioambientales y la legislación europea. A continuación, el capítulo cinco el potencial de la Inteligencia

Artificial destaca las ventajas y el poder transformador de esta tecnología en áreas clave. Finalmente, el capítulo conclusión sintetiza los hallazgos más relevantes del trabajo y ofrece una reflexión final sobre el futuro de la inteligencia artificial en la sociedad, cerrando el estudio con una visión integral de sus retos y sus posibilidades.

2. Inteligencia artificial: bases teóricas

2.1 Historia y evolución de la inteligencia artificial

La historia de la inteligencia artificial IA es un proceso complejo y multidisciplinar que está profundamente vinculado a los avances en computación, lógica y teorías de la mente. Sus raíces formales comienzan a mediados del siglo XX, cuando Alan Turing planteó la pregunta fundamental: ¿pueden las máquinas pensar? Para responder a esto diseñó el test de Turing, un criterio para evaluar si una máquina puede exhibir un comportamiento indistinguible del ser humano, asentando así las bases conceptuales para el desarrollo de la IA moderna (Turing, 1950; Yébenes, 2024).

Durante las primeras décadas de la inteligencia artificial, uno de los hitos fundamentales es el desarrollo de las redes neuronales artificiales, cuyo origen se remonta a 1943 con el trabajo Walter Pitts y Warren McCulloch. Estos autores propusieron un modelo matemático simplificado inspirado en la neurobiología que conceptualizaba las neuronas como unidades binarias que se activan o no en función de ciertos umbrales. Su modelo formalizó cómo las neuronas podrían conectarse para procesar información estableciendo un puente entre la biología y la computación. Este enfoque pionero permitió sentar las bases para la representación matemática y computacional de los procesos cognitivos y la idea de que una red de unidades simples podía en conjunto realizar tareas complejas (Mcculloch & Pitts, 1943).

En 1956, John McCarthy acuñó el término inteligencia artificial durante la conferencia de Dartmouth, considerada como el nacimiento oficial del campo. Por lo tanto, se promovió la idea de que las máquinas podían ser programadas para realizar tareas que requieren inteligencia humana, dando inicio a un intenso periodo de investigación y desarrollo. Durante la llamada “Edad dorada de la IA”, entre 1956 y 1974, se desarrollaron programas capaces de resolver problemas lógicos, hasta entrenar a los algoritmos para jugar al ajedrez, aunque su eficacia estaba limitada a entornos muy controlados (Abeliuk & Gutiérrez, 2021).

A partir del trabajo preliminar de Pitts y McCulloch, en 1958 Frank Rosenblatt desarrolló el Perceptrón, el primer algoritmo de red neuronal capaz de aprender mediante un proceso supervisado. El Perceptrón es un modelo simple que ajusta los pesos internos de la neurona para clasificar datos en diferentes categorías. Rosenblatt demostró que el modelo podía aprender a reconocer patrones visuales básicos, lo que despertó un gran interés en la comunidad científica por la posibilidad de que las máquinas

aprendieran de datos y no solo siguieran reglas preprogramadas. Sin embargo, *El Perceptrón* también evidenció ciertas limitaciones, como que era incapaz de resolver problemas no linealmente separables, como la función XOR, lo que llevó a una crítica significativa del modelo en la década de 1960 y a un estancamiento temporal en el campo (Rosenblatt, 1958).

Estas críticas fueron reforzadas por el influyente libro *Perceptrons* de Marvin Minsky y Seymour Papert en 1969, dónde señalaron las limitaciones teóricas y prácticas de las primeras redes neuronales, lo que provocó un periodo de menor interés y financiación conocido como "el primer invierno de la IA" (Minsky & Papert, 1969). Sin embargo, las ideas de Pitts, McCulloch y Rosenblatt no quedaron obsoletas, con el tiempo, gracias a las mejoras en algoritmos de aprendizaje como la del retropropagación y el aumento de capacidad computacional. Las redes resurgen con fuerza en la era de aprendizaje profundo convirtiéndose así en una de las tecnologías más importantes para el procesamiento de datos complejos en la era contemporánea.

ELIZA, creado por Joseph Weizenbaum en 1966, constituye uno de los primeros sistemas de procesamiento del lenguaje natural diseñados para simular una conversación con un ser humano. Mediante un conjunto de reglas de reescritura y patrones lingüísticos predefinidos, ELIZA era capaz de generar respuestas simulando una interacción conversacional con el usuario particularmente bajo el guion de un psicoterapeuta. Aunque no comprendía semánticamente los enunciados, logró demostrar cómo estructuras sintácticas simples podían introducir la apariencia de comprensión en un sistema artificial (Weizenbaum, 1966)

Los sistemas expertos, una de las primeras aplicaciones exitosas de la inteligencia artificial, surgió en los años 60, pero alcanzaron un mayor auge en la década de 1980, cuando comenzaron a implementarse de forma efectiva en sectores como la medicina, la ingeniería y las finanzas. Estos sistemas simulan el razonamiento humano mediante reglas y bases de conocimiento permitiendo resolver problemas complejos en contextos específicos. Edward Feigenbaum, considerado como uno de los pioneros en este campo describió esta disciplina como "Ingeniería del conocimiento" subrayando la importancia de formalizar el saber experto para su aplicación computacional (Feigenbaum, 1984).

A lo largo de su historia, la IA ha atravesado varios "inviernos", es decir, periodos de estancamiento causados por expectativas tecnológicas no cumplidas y la consiguiente reducción de la financiación. Estos momentos críticos no frenaron la investigación, sino que impulsaron el perfeccionamiento de algoritmos y bases conceptuales, que serían cruciales para futuros avances (Haenlein & Kaplan, 2019).

El resurgimiento decisivo de la inteligencia artificial llegó a finales del siglo XX y principios del siglo XXI, gracias a la confluencia de varios factores clave. Algunos hitos como la victoria de Deep blue de IBM sobre el campeón mundial de ajedrez Garry Kaspárov en 1997 (IBM, 1997) o el éxito de IBM Watson en Jeopardy! en 2011 (IBM, 2011), demostraron un enorme progreso en la capacidad de procesamiento de búsqueda y comprensión del lenguaje natural.

En el siglo XXI la IA tuvo un gran avance gracias al aumento exponencial de la capacidad computacional (GPUs), la disponibilidad masiva de los datos y la introducción de algoritmos de aprendizaje profundo. Entre estos avances, destaca el modelo Transformer, ha presentado en 2017 en el Paper “Attention Is All You Need”, que revolucionó el procesamiento del lenguaje natural al utilizar mecanismos de atención para captar relaciones contextuales en datos secuenciales sin depender de estructuras recurrentes (Vaswani et al., 2017). Este modelo ha mejorado significativamente la eficiencia y la precisión de los sistemas, como traductores automáticos y los asistentes virtuales, allanando el camino para la IA generativa que alcanzó una conciencia pública masiva con el lanzamiento de ChatGPT por OpenAI en 2022 (OpenAI, 2022).

Con todos estos avances, la IA ya ha dejado ser únicamente un intento de imitar la inteligencia humana para convertirse en una infraestructura cognitiva que transforma en procesos sociales, económicos y comunicativos. Esta tecnología introduce nuevas formas de actuar en la toma de decisiones, la producción de conocimiento y la mediación simbólica en las sociedades contemporáneas (Dick, 2019).

2.2 Definición de inteligencia artificial

La inteligencia artificial (IA) es un concepto amplio y en constante evolución, lo que dificulta establecer una única definición precisa. En términos generales, se trata de una rama de la informática cuyo propósito es diseñar tecnologías capaces de reproducir capacidades asociadas a la inteligencia humana, como el aprendizaje, la toma de decisiones, la resolución de problemas o la capacidad de autocorrección.

Mediante el desarrollo de algoritmos avanzados y sistemas especializados, los dispositivos dotados de IA pueden ejecutar tareas cognitivas de forma autónoma. Estas tecnologías son capaces de percibir su entorno, identificar objetos y patrones, interpretar y generar un lenguaje natural, aprender de la experiencia previa, ofrecer recomendaciones personalizadas e incluso actuar sin intervención humana directa.

A diferencia de la visión reduccionista del pasado, en la actualidad la inteligencia artificial no tiene por qué ser entendida como un sustituto del ser humano, sino como una

herramienta que puede complementar y potenciar sus capacidades. Sus aplicaciones, cada vez más extendidas, tienen como finalidad optimizar procesos, aumentar la eficiencia y mejorar la interacción entre usuarios y sistemas tecnológicos.

Definir con precisión qué es la inteligencia artificial resulta complejo, ya que sus límites y capacidades evolucionan constantemente. Algunos autores la definen como "la ciencia y la ingeniería de las máquinas con capacidades que se consideran inteligentes según los estándares de la inteligencia humana", destacando así el vínculo directo entre la IA y la imitación del comportamiento humano (Boden et al., 2017).

2.3 Tipos de Inteligencia Artificial

Dada la amplitud del campo de la inteligencia artificial, existen diversas formas de clasificarla. A continuación, se presentan las más reconocidas, que permiten comprender mejor su evolución, funcionamiento y niveles de complejidad.

2.3.1 Clasificación por el nivel de inteligencia

Esta clasificación es la ampliamente aceptada, se basa en la capacidad de un sistema de IA para igualar o superar la inteligencia humana. Se distinguen tres tipos:

- **IA débil o estrecha (ANI)**: Sistemas diseñados para realizar tareas específicas. No tienen comprensión ni conciencia, su toma de decisiones está limitada a su programación. Ejemplos de este tipo son los asistentes virtuales (Alexa, Siri…), sistemas de reconocimiento facial y plataformas de recomendación de contenido. Se trata de la forma más común y extendida de IA en la actualidad.
- **IA general (AGI):** Este tipo incluye sistemas con capacidad de realizar cualquier tarea intelectual que pueda hacer un ser humano, con razonamiento, autonomía, adaptabilidad y aprendizaje. Actualmente, se encuentran en fase de investigación y no hay aplicaciones reales disponibles.
- **IA superinteligente (ASI)**: Este tipo de inteligencia artificial representa el nivel más avanzado dentro del desarrollo teórico de la IA. Hace referencia a sistemas que superarían la inteligencia humana en múltiples aspectos, desde la resolución de problemas complejos hasta la creatividad. Serían capaces de realizar las tareas con un nivel de eficacia y autonomía muy superior al del ser humano. El objetivo final del desarrollo de la inteligencia artificial es poder alcanzar a esta categoría. Sin embargo, su desarrollo plantea importantes desafíos éticos, legales y sociales que requerirían una profunda reflexión y regulación (Kingler, 2024).

2.3.2 Clasificación de Arentz Hintze

Arend Hintze, profesor de Biología Integrada y Ciencias de la Computación en la Universidad de Michigan. La clasificación que él propone se basa en la capacidad predictiva, aprendizaje y toma de decisiones, en los diferentes niveles de complejidad que puede alcanzar una máquina. Divide la inteligencia artificial en cuatro tipos, ordenados según su nivel de desarrollo y autonomía:

- **Máquinas reactivas:** Este tipo se asemeja a la inteligencia artificial estrecha, ya que incluye los sistemas más básicos diseñados para ejecutar tareas concretas con unas normas predefinidas. No poseen capacidad de toma de decisión, porque no almacenan memoria, experiencias ni recuerdos. Su funcionamiento se limita al reconocimiento de patrones específicos predeterminados sin capacidad de adaptación al medio o al contexto. Un ejemplo emblemático es la supercomputadora Deep blue de IBM, que logró vencer a Kaspárov en una partida de ajedrez a finales de los 90. Este sistema fue diseñado exclusivamente para jugar al ajedrez, y por eso pudo ejecutar esa tarea de forma muy precisa, reconociendo las piezas, sus movimientos y anticipando jugadas (IBM, 1997). Otros ejemplos actuales son los sistemas de gestión de redes, robots industriales o líneas de producción automatizadas,
- **Memoria limitada:** En este tipo se incluyen sistemas con capacidad concreta de almacenamiento y procesamiento de datos pasados. Esto les permite aprender a partir de datos y experiencias previas, aunque de forma limitada y especifica. Muchos de los sistemas actuales pertenecen a esta categoría, como los coches autónomos que pueden adaptarse a las normas de circulación de cada tramo y pueden tomar las decisiones en tiempo real, o los asistentes virtuales, que responden según el historial de interacción con el usuario.
- **Teoría de la mente:** este tipo de inteligencia artificial aún se encuentra en una etapa de desarrollo. Se refiere a los sistemas que serían capaces de interpretar y comprender las emociones, intenciones, creencias y expectativas de los seres humanos. El objetivo es que las máquinas consigan interactuar socialmente de manera similar a como lo haría un ser humano, teniendo en cuenta no solo en el lenguaje verbal, sino el contexto emocional y psicológico en el que se encuentra. Aunque aún no existen aplicaciones completamente funcionales con estas capacidades, su desarrollo está vinculado a avances en campos como la robótica social, la psicología cognitiva y las neurociencias. Se considera una etapa intermedia entre la inteligencia artificial general (AGI) y la superinteligente

(ASI), representando un paso crucial hacia una IA con capacidades sociales y comunicativas complejas.

- **Autoconciencia**: es el nivel más avanzado que existe, comparable a la superinteligencia (ASI). Este tipo las máquinas no solo comprenderían las emociones y las necesidades humanas, sino que desarrollarían una autoconciencia. Es decir, serían capaces de entender su propia existencia, reconocer su identidad y tomar decisiones basadas en esa autopercepción. El desarrollo de una IA autoconsciente implicaría que las máquinas pudieran tener una comprensión profunda de su entorno, interactuar con él de manera autónoma y evolucionar de acuerdo con sus experiencias y objetivos. A pesar de que, actualmente no existen sistemas autoconscientes, este tipo de inteligencia artificial representa un tema fundamental en debates éticos, así como sobre el futuro de la tecnología. El reto principal reside en determinar cómo se podría lograr que una máquina tuviera conciencia de sí misma, así como las implicaciones que esto tendría para la sociedad (Hintze, 2016).

2.3.3 Clasificación de Stuart Russell y Peter Norvig

Stuart Russell y Peter Norvig, dos destacados investigadores en el campo de la inteligencia artificial proponen una clasificación en su obra "*Inteligencia Artificial: Un Enfoque Moderno*". Su clasificación se enfoca en las diferentes formas en que una máquina puede ser diseñada para resolver problemas y tomar decisiones en base al comportamiento y la racionalidad. Esta clasificación se divide en cuatro categorías:

- **Sistemas que piensan como humanos**: Las máquinas tienen como finalidad replicar el proceso del pensamiento humano. No se busca únicamente que imiten el comportamiento humano, sino también que resuelvan problemas de una manera similar a cómo lo haría un ser humano. Estos sistemas deben ser capaces de razonar, aprender y adaptarse a nuevas situaciones, basándose en procesos cognitivos que emulan el pensamiento humano. Un ejemplo de este tipo, son las redes neuronales artificiales que están inspiradas en la estructura del cerebro humano y se utilizan para automatizar tareas complejas, como el reconocimiento de patrones, la clasificación de datos o la predicción de resultados.
- **Sistemas que actúan como humanos**: Se refiere a máquinas diseñadas para realizar tareas de forma similar a los humanos, pero de manera más eficiente. El objetivo no es imitar el proceso de pensamiento humano, sino replicar su comportamiento. Un ejemplo representativo son los robots, que pueden realizar

actividades humanas como caminar, manipular objetos o interactuar en entornos sociales, sin necesariamente pensar como una persona.

- **Sistemas que piensan racionalmente:** Engloba los sistemas cuyo objetivo es imitar el razonamiento lógico y racional. A través del uso de modelos computacionales, se analizan las capacidades mentales para diseñar sistemas que sean capaces de percibir su entorno, razonar de forma lógica y actuar en consecuencia. Un ejemplo de esta categoría son los sistemas expertos, los cuales simulan el proceso de toma de decisiones de los humanos basándose en el conocimiento y la experiencia aplicada.
- **Sistemas que actúan racionalmente:** Esta categoría agrupa a los sistemas cuyo propósito es actuar de la manera más racional posible, comparable al comportamiento humano, tomando las decisiones más lógicas y eficientes posibles para alcanzar un objetivo. Los agentes inteligentes, como los utilizados en sistemas de navegación autónoma o en algoritmos de optimización, son ejemplos de este tipo de inteligencia artificial (Russell & Norvig, 2010).

En conjunto, estas clasificaciones ofrecen una visión clara sobre como la inteligencia artificial está organizada, abarcando desde sus aplicaciones actuales hasta los posibles desarrollos futuros. A medida que la investigación en este campo continúa, estas clasificaciones serán clave para orientar la implementación de la tecnología, permitiendo una mejor comprensión de sus capacidades, limitaciones y el impacto que tendrá en diversas áreas. Además, nos ayudarán a anticipar los desafíos éticos, sociales y técnicos asociados, asegurando que se tomen decisiones informadas sobre su uso y evolución.

2.4 Principios de funcionamiento de la inteligencia artificial

El funcionamiento de la inteligencia artificial (IA) se basa en un conjunto de principios tecnológicos y matemáticos que permiten a las máquinas simular funciones cognitivas humanas, como el aprendizaje, la percepción y la toma de decisiones. A través de la combinación de conocimiento procedentes de principios que integran disciplinas como la estadística, las matemáticas, la informática y la neurociencia, es posible desarrollar sistemas capaces de mejorar su desempeño a través de la experiencia y adaptarse a entornos dinámicos.

En este apartado, se explorarán los principales fundamentos operativos que hacen posible el desarrollo y el funcionamiento de la inteligencia artificial. Se abordarán, en

primer lugar, el aprendizaje automático (Machine learning), seguido del aprendizaje profundo (Deep learning), un campo que ha potenciado los avances más recientes. También se explicarán las redes neuronales artificiales, pilares esenciales de la IA contemporánea, y finalmente se examinará el papel emergente de la inteligencia artificial generativa.

Cada una de estas técnicas representan un paso fundamental en la evolución de la IA moderna y serán tratadas con el fin de ofrecer una visión integral de sus principios operativos, aplicaciones y desafíos actuales.

2.4.1 Aprendizaje automático

El aprendizaje automático representa una de las ramas más influyentes y dinámicas de la inteligencia artificial. Su principal característica es permitir que los sistemas informáticos aprendan directamente los datos, sin requerir una programación explícita para cada tarea específica. Este enfoque ha revolucionado múltiples sectores al dotar a las máquinas de capacidad de detectar patrones complejos, realizar predicciones precisas y tomar decisiones autónomas en contextos variables (Alpaydin, 2016).

A medida que la digitalización masiva ha dado lugar a una explosión de datos, lo que conocemos como Big Data, el machine learning ha emergido como una herramienta clave para extraer valor de esta información. La capacidad de los algoritmos para construir modelos predictivos o descriptivos se ha convertido en el motor de innovación en ámbitos tan diversos, como la salud, la seguridad, el comercio electrónico o la ingeniería (Alpaydin, 2016).

Uno de los grandes atractivos del aprendizaje automático es su habilidad para generar programas automáticamente a partir de los datos. Para que este proceso sea exitoso, es fundamental comprender tres componentes esenciales: la representación (cómo se modelan los datos), la evaluación (cómo se mide el rendimiento del modelo) y la optimización (cómo se mejora el modelo a lo largo del tiempo). Asimismo, la ingeniería de características y la gestión de sobreajuste (overfitting) son aspectos cruciales para garantizar que los modelos se generalicen correctamente a nuevos conjuntos de datos (Domingos, 2012).

Entre los enfoques más comunes del machine learning se encuentran:

- Aprendizaje supervisado: el modelo se entrena con datos etiquetados para predecir resultados futuros, siendo la clasificación una de sus técnicas clave.

- Aprendizaje no supervisado: trabaja con datos no etiquetados para descubrir patrones ocultos o estructuras subyacentes, como ocurre en el análisis de clústeres.
- Aprendizaje por refuerzo: un agente autónomo aprende a través de la interacción con su entorno, maximizando recompensas y minimizando penalizaciones mediante prueba y error.

Estas categorías permiten abordar una gran variedad de problemas y son fundamentales para el desarrollo de sistemas inteligentes capaces de adaptarse a situaciones complejas (Alnuaimi & Albaldawi, 2024).

El machine learning se apoya en diversos algoritmos para aprender de los datos y realizar tareas específicas. Dentro del aprendizaje supervisado, donde se entrena con datos etiquetados para predecir un resultado, encontramos la regresión lineal para predecir valores continuos, la regresión logística para dar clasificación binaria con dos categorías: los árboles de decisión, que modelan decisiones como un diagrama de flujo, y las Máquinas de Vectores de Soporte, que buscan el mejor hiperplano para separar clases. Para el aprendizaje no supervisado, dónde se buscan patrones en datos sin etiquetar, el algoritmo K-Means es fundamental para agrupar puntos de datos en clústeres. Finalmente, el aprendizaje por refuerzo, el algoritmo Q-Learning permite a un agente aprender las mejores acciones en un entorno para maximizar recompensas a largo plazo, todo ello a través de prueba y error (Alnuaimi & Albaldawi, 2024).

A pesar de estos desafíos, el aprendizaje automático continúa evolucionado y actualizándose rápidamente, con avances que buscan superar sus limitaciones. Una dirección significativa es el desarrollo del aprendizaje automático automatizado (AutoML), que busca automatizar gran parte del proceso de desarrollo de modelos de Machine learning, incluyendo la selección de algoritmos y la optimización de hiperparámetros (Balaji & Allen, 2018).

Desde una visión más integral, el desarrollo del aprendizaje automático puede entenderse como una evolución tecnológica continua, desde los grandes ordenadores hasta los dispositivos móviles actuales. Este recorrido ha estado acompañado por avances teóricos y prácticos, como el uso de algoritmos de reconocimiento de patrones, el diseño de sistemas de recomendación y la aparición de redes neuronales artificiales inspiradas en el funcionamiento del cerebro humano (Alpaydin, 2016).

2.4.2 Aprendizaje profundo

El aprendizaje profundo o Deep learning es una subrama del machine learning, basado en redes neuronales artificiales con múltiples capas que permiten aprender representaciones jerárquicas de los datos. Estas múltiples capas transforman la información de manera progresiva, extrayendo características desde niveles simples hasta conceptos altamente abstractos, lo que facilita el reconocimiento de patrones complejos en grandes volúmenes de datos (Schmidhuber, 2015). Este enfoque ha demostrado ser muy eficaz en campos dónde la complejidad y heterogeneidad de los datos requieren modelos capaces de capturar relaciones lineales profundas (Lecun et al., 2015).

2.4.2.1 Redes neuronales artificiales

Las redes neuronales artificiales (RNA) son modelos computacionales inspirados en la estructura del cerebro humano, diseñado para reconocer patrones y resolver problemas complejos mediante la interconexión de unidades simples llamadas neuronas artificiales (Mcculloch & Pitts, 1943). En la Figura 1, se muestra el parecido entre una neurona humana y una neurona artificial señalando cada una de sus partes.

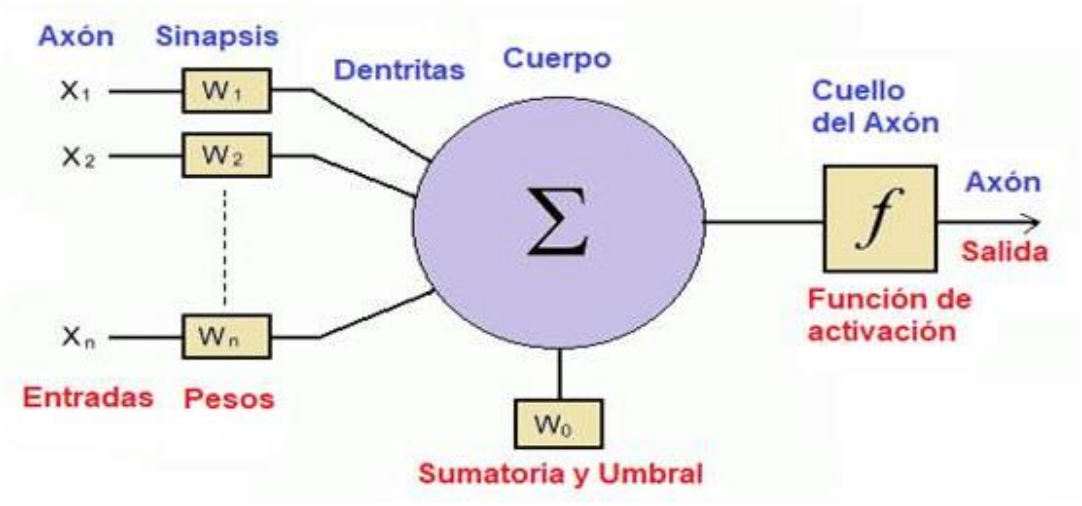

Figura 1. Comparación de una neurona humana con una neurona artificial

Fuente: (Sarmiento-Ramos, 2020; T. Hagan et al., 2014)

Estas redes están formadas por capas de nodos que transforman progresivamente los datos de entrada hasta generar una salida, siguiendo un enfoque jerárquico de representación de la información (Minsky & Papert, 1969).

Una red neuronal típica, como la que aparece en la Figura 2, consta de tres tipos de capas: la capa de entrada, que recibe los datos iniciales, una o más capas ocultas, que realizan las transformaciones, y la capa de salida, que produce la

respuesta final del modelo. Cada neurona procesa sus entradas mediante una combinación lineal ponderada seguida de una función de activación no lineal, lo que permite capturar relaciones complejas entre variables (Dias et al., 2004).

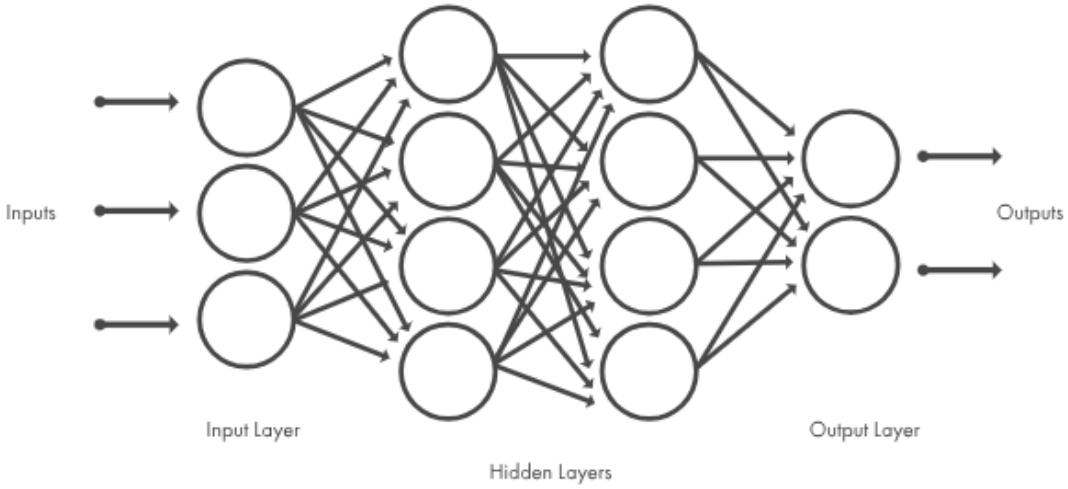

Figura 2. Estructura de una red neuronal

Fuente: (The MathWorks, 2025)

El aprendizaje de una red neuronal se basa en un algoritmo de optimización que ajusta los pesos sinápticos de las conexiones. Uno de los métodos más utilizados es el algoritmo de retropropagación (backpropagation), el cual calcula el error entre la salida producida por la red y la salida deseada, y lo difunde hacia atrás a lo largo de las capas para ajustar los pesos en función de ese error. Este proceso se repite iterativamente hasta que se minimiza la función de pérdida mediante técnicas como el descenso del gradiente (Lecun et al., 2015; Schmidhuber, 2015).

Las redes neuronales pueden clasificarse en distintas arquitecturas dependiendo del tipo de datos que procesan. Por ejemplo, las redes convoluciones (CNN) son especialmente eficaces para datos espaciales como imágenes, mientras que las redes recurrentes (RNN) son adecuadas para datos secuenciales como texto o series temporales, ya que pueden conservar información de estados anteriores (Misra & Saha, 2010).

Estas arquitecturas han demostrado ser muy efectivas en tareas como el reconocimiento de voz, la traducción automática o la visión artificial. Sin embargo, las redes neuronales también presentan importantes limitaciones. Entre ellas, se encuentra la opacidad de los algoritmos de “caja negra”, ya que a menudo es difícil interpretar como llegan sus decisiones, lo que limita su uso en contextos que exigen transparencia y explicabilidad (Lecun et al., 2015).

Además, su entrenamiento requiere una enorme cantidad de datos etiquetados, lo que lleva mucho tiempo y gran potencia computacional que no siempre está disponible. También pueden ser vulnerables al sobreajuste, especialmente cuando se entrenan con conjuntos de datos insuficientes o no representativos (Lecun et al., 2015; Schmidhuber, 2015). Otro desafío relevante es su sensibilidad a las perturbaciones adversas, que son pequeñas modificaciones en los datos de entrada que pueden alterar significativamente la salida del modelo, comprometiendo su robustez (Lecun et al., 2015) .

2.4.2.2 Inteligencia artificial generativa

La inteligencia artificial generativa (IAG) se refiere a un conjunto de técnicas de aprendizaje profundo capaces de generar contenido nuevo y original a partir de grandes volúmenes de datos ya existentes. A diferencia de los enfoques anteriores discriminativos que clasifican y predicen etiquetas, los modelos generativos aprenden la distribución subyacente de los datos y la utilizan para producir nuevas soluciones que conservan la coherencia con los patrones observados durante el entrenamiento (Bengesi et al., 2024).

Entre las arquitecturas más destacadas, se encuentran las redes generativas antagónicas (GAN), introducidas por Ian Goodfellow, que funcionan a través de un juego competitivo entre dos redes: un generador, que crea muestras falsas, y un discriminador, que intenta distinguir entre muestras reales y falsas (Goodfellow et al., 2014). Esta arquitectura ha sido ampliamente utilizada en la generación realista de imágenes, deepfakes y síntesis de audios. Otro ejemplo son los Autoencoder Variacionales (VAE), que codifican los datos en una representación latente para reconstruirlos, permitiendo la manipulación y generación controlada de contenido (Bengesi et al., 2024).

Los transformadores, basados en el mecanismo propuesto en el Paper “Attention Is All You Need”, revolucionaron la IA generativa al mejorar la capacidad de modelar dependencias a largo plazo en secuencias (Vaswani et al., 2017). Este avance posibilita la aparición de modelos como GPT (Generative Pre-trained Transformer), que pueden generar textos coherentes, traducir idiomas o responder preguntas con fluidez casi humana. Los modelos de difusión, por su parte, son técnicas emergentes que crean imágenes partiendo del ruido aleatorio refinándolo paso a paso, hasta producir resultados visuales de alta calidad (Bengesi et al., 2024).

Las aplicaciones de IA generativa son amplias y van desde la creación de obras artísticas, videojuegos y materiales educativos, hasta el diseño molecular en biomedicina o la generación de prototipos de ingeniería. También se ha utilizado en la síntesis de datos para mejorar la privacidad o aumentar la diversidad en conjuntos de entrenamiento. No obstante, estudios recientes advierten que su adopción puede tener efectos ambivalentes en contextos sociales, por un lado, permite democratizar el acceso a herramientas creativas y automatizar procesos complejos, pero por otro lado, puede ampliar brechas socioeconómicas, si su acceso y beneficios no se distribuyen equitativamente (Capraro et al., 2024).

La inversión global en inteligencia artificial generativa ha experimentado un crecimiento sostenido en los últimos años, con un aumento muy notable a partir de 2022, como muestra la Figura 3 (Our World in Data, 2024). Esto coincide con la introducción al mercado de ChatGPT de IBM (OpenAI, 2022). Esta tendencia refleja el creciente interés por parte de gobiernos, empresas y centros de investigación en el desarrollo y aplicación de estas tecnologías.

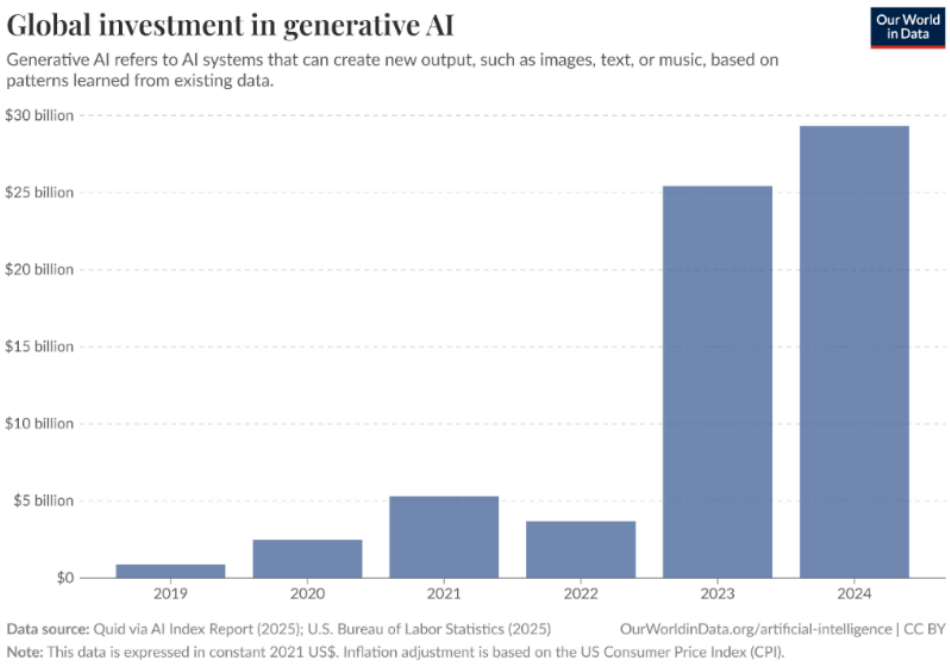

Figura 3. Gráfico de barras inversión global en IA generativa (2019-2024)

Fuente:(Our World in Data, 2024)

No obstante, esta tecnología presenta algunas limitaciones. Uno de los principales desafíos es el control sobre la salida generada, los modelos pueden producir resultados inesperados, incoherentes o incluso incorrectos. Además, la IAG puede amplificar sesgos presentes en los datos de entrenamiento,

perpetuando estereotipos o generando contenido discriminatorio. Otro problema es la opacidad, ya que entender porque un modelo generó determinada respuesta puede ser muy difícil. Otra preocupación ética es el posible uso malicioso de estos sistemas para crear desinformación, suplantación de identidad o contenido manipulado (Sengar et al., 2024).

3. Impacto social de la Inteligencia Artificial

La expansión de la inteligencia artificial en múltiples ámbitos de la vida cotidiana está generando profundos cambios en las dinámicas sociales, culturales y económicas. Su integración en sectores como la educación, la salud, la movilidad, el entretenimiento o el mundo laboral no solo está redefiniendo procesos y prácticas, sino que también plantea nuevos desafíos en términos de equidad, acceso, autonomía y calidad de vida.

Las percepciones sociales sobre el impacto de la inteligencia artificial también juegan un papel clave en su desarrollo y aceptación. Tal como se muestra en la Figura 4, existen diferencias notables según edad, género y nivel socioeconómico respecto a si la IA ayudará o perjudicará a las personas en los próximos 20 años (Our World in Data, 2024).

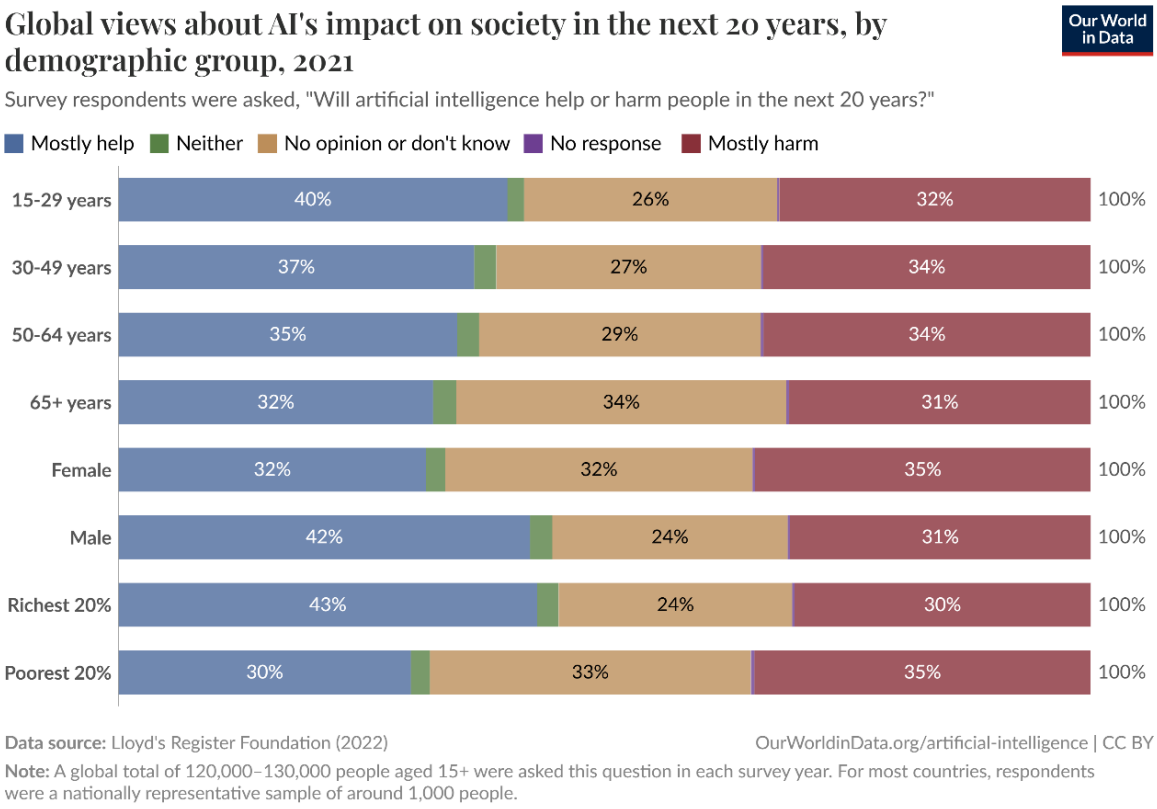

Figura 4. Percepciones globales sobre el impacto de la IA en los próximos 20 años, según grupos demográficos.

Fuente: (Our World in Data, 2024)

A diferencia de los riesgos éticos causados por la IA, tratados en apartados específicos, esta sección se centrará en el impacto directo de la IA sobre los comportamientos sociales, las relaciones humanas y las estructuras institucionales. La inteligencia artificial al involucrarse cada vez más en las interacciones entre individuos y sistemas

moldea decisiones, filtra información, automatiza tareas y condiciona la participación social. Por ello, resulta fundamental analizar sus efectos en sectores clave, valorando no solo las oportunidades que ofrece, sino también los desafíos y tensiones que genera.

Este análisis abarca desde el ámbito económico hasta el entretenimiento, pasando por el empleo, la salud, la educación y el transporte, con el objetivo de ofrecer una visión crítica, pero equilibrada sobre como la IA está reformulando la vida en sociedad.

3.1 Economía y finanzas

En los últimos años, la inteligencia artificial ha pasado de ser una herramienta emergente para convertirse en un componente estratégico fundamental dentro de la estructura económica y financiera global. Más allá de su aplicación técnica, la IA está reconfigurando las dinámicas de desarrollo económico, la productividad empresarial y la estructura del empleo generando impactos sociales de gran alcance.

Desde una perspectiva macroeconómica, la IA está contribuyendo a una mayor eficiencia en la asignación de recursos, al fortalecimiento de la innovación y al impulso del crecimiento económico sostenido. Su capacidad para optimizar procesos, anticipar tendencias y reducir incertidumbres, para convertir esta tecnología en un catalizador clave de desarrollo productivo, especialmente en economías que buscan modernizar su infraestructura tecnológica e industrial (Trabelsi, 2024). La adopción de tecnologías basadas en inteligencia artificial está creciendo de forma significativa en distintos sectores y regiones del mundo. Como se observa en la figura 5, el porcentaje de empresas que implementan soluciones de IA, en al menos una función de negocio, ha aumentado de manera notable entre 2021 y 2024, especialmente en regiones como América del Norte y Europa donde supera ya el 80% (Our World in Data, 2024).

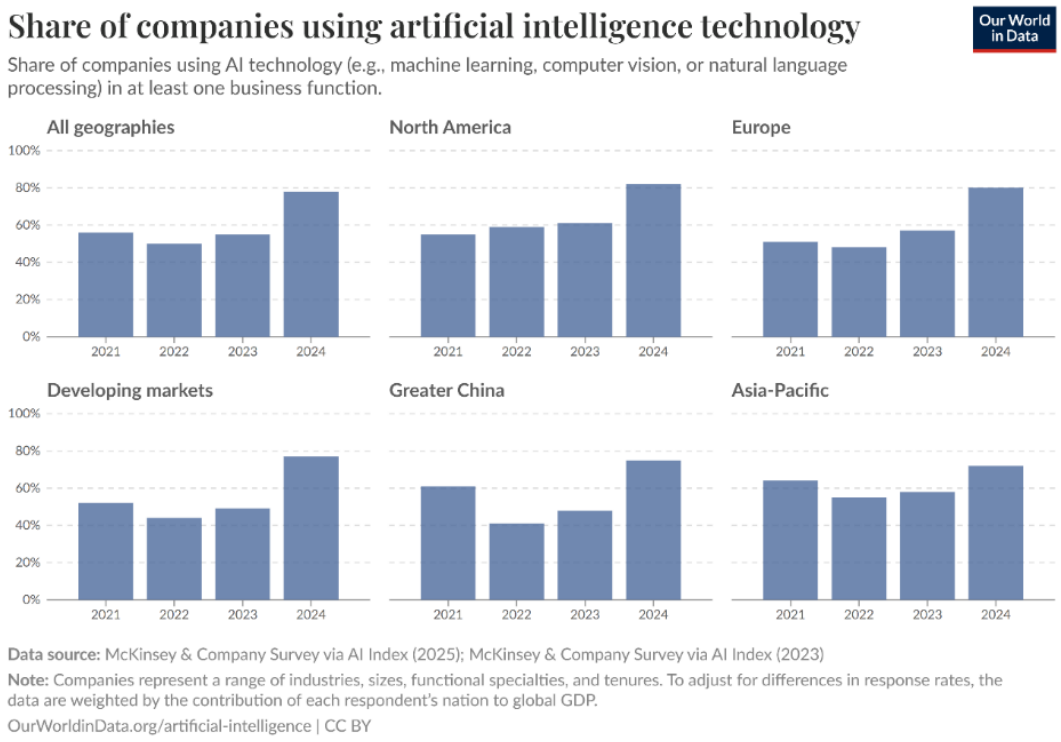

Figura 5. Porcentaje de empresas que utilizan tecnología de inteligencia artificial en diferentes regiones del mundo (2021–2024)

Fuente: (Our World in Data, 2024)

Paralelamente, el impacto de la inteligencia artificial en el bienestar económico ha sido notable en generar ganancias, incrementar la productividad total de los factores y aumentar la rentabilidad de las inversiones públicas y privadas. No obstante, este impacto se distribuye de forma desigual entre países, regiones y sectores, lo que podría acentuar las brechas de desarrollo, si no es acompañada de políticas de inclusión digital. La concentración de capacidades tecnológicas en mercados desarrollados tiende a favorecer a grandes corporaciones, mientras que los países con menor infraestructura digital enfrentan riesgos de marginación tecnológica (Lu, 2021).

En el ámbito financiero, la IA ha transformado profundamente las prácticas tradicionales. Se ha integrado en procesos como la auditoría, la contabilidad automatizada, la gestión de riesgos y la evaluación de crédito, lo que ha permitido una mejora significativa en la precisión de los sistemas contables y una reducción de errores humanos (Sánchez-Caguana et al., 2024). Su implementación ha impulsado el desarrollo de servicios financieros personalizados, eficientes y seguros, tanto en la banca como en el sector de las aseguradoras (Alcázar-Blanco et al., 2024).

La capacidad de la IA para procesar datos en tiempo real y reconocer patrones complejos ha revolucionado la toma de decisiones financieras. Las organizaciones la

emplean para prever crisis económicas, detectar fraudes, gestionar inversiones y formular estrategias comerciales de alta precisión. Sin embargo, este entorno híper automatizado exige nuevos marcos regulatorios que garanticen la transparencia algorítmica y la gobernanza ética de los datos financieros (Bahoo et al., 2024).

La inteligencia artificial ha revolucionado el trading financiero a través del uso de algoritmos de aprendizaje automático y redes neuronales profundas, que permiten ejecutar operaciones algorítmicas de alta frecuencia. Estos modelos analizan flujos de datos históricos y en tiempo real, como precios, volúmenes, noticias financieras o señales de mercado, para identificar patrones y predecir movimientos bursátiles con una velocidad y precisión inalcanzables para los humanos. Algunas técnicas como el aprendizaje por refuerzo están siendo aplicadas para optimizar estrategias adaptativas que ajustan sus decisiones en función del entorno dinámico del mercado (Alcázar-Blanco et al., 2024; Bahoo et al., 2024)

Además, la inteligencia artificial ha conseguido desempeñar un papel muy protagonista en la reconfiguración del mercado laboral. Aunque ha generado empleos altamente cualificados en tecnologías emergentes, también ha provocado desplazamientos laborales en sectores que tradicionalmente dependían de mano de obra rutinaria y tareas repetitivas. Esto plantea el desafío de adaptar la educación y la formación profesional a las nuevas exigencias del mercado. La transición digital si no se gestiona adecuadamente puedan acentuar la polarización del empleo y aumentar las desigualdades económicas (Pinto Molina, 2023).

3.2 Empleo y automatización

La inteligencia artificial está transformando no solo la estructura del mercado laboral, sino también el bienestar psicológico y emocional de los trabajadores. Este impacto es doble, por un lado, redefine qué tareas realizan los humanos y qué tareas delegan a las máquinas; por otro, altera las condiciones laborales y el equilibrio entre la vida profesional y personal.

Lejos de ser un fenómeno homogéneo, el impacto de la IA varía según el tipo de tarea, el nivel de ingresos de los países, el capital humano disponible y las políticas de adaptación implementadas.

Uno de los aspectos más debatidos es si la IA está eliminando empleos o simplemente modificando las tareas dentro de los mismos. La evidencia sugiere que la IA tiende a automatizar tareas específicas, particularmente aquellas rutinarias, repetitivas o

altamente estructuradas, más que a sustituir ocupaciones enteras. En este sentido lo que se está produciendo es una redistribución funcional del trabajo, donde las habilidades humanas complementan a los sistemas automatizados en lugar de competir directamente con ellos (Bonney et al., 2024).

Este fenómeno genera tensiones significativas en el mercado laboral, ya que beneficia principalmente a los trabajadores con competencias digitales y capacidad de adaptación, mientras que expone a mayor vulnerabilidad a quienes ocupan puestos menos cualificados. La IA está contribuyendo a una creciente segmentación laboral, ampliando la brecha entre empleos altamente especializados y aquellos fácilmente automatizables, lo cual incrementa la desigualdad dentro del mismo país y entre países (Freire, 2025).

Además de las consecuencias estructurales, la IA tiene un fuerte impacto en el bienestar subjetivo de los empleados. Según el modelo de demandas laborales y de recursos, el efecto de estas tecnologías puede ser ambivalente: cuando permiten reducir la sobrecarga o facilitar la organización del trabajo pueden aumentar la satisfacción laboral, sin embargo, cuando generan vigilancia constante, presión por productividad o temor al reemplazo, provocan ansiedad, desmotivación o desconexión emocional (Chuang et al., 2025).

Este efecto se amplifica en contextos donde la transformación digital avanza más rápido que la reconversión profesional. En varias economías europeas, por ejemplo, el uso intensivo de tecnologías de la información ha provocado desplazamientos laborales a corto plazo, afectando sobre todo a trabajadores de baja cualificación que no han podido adaptarse a las nuevas demandas del mercado. Aunque en el largo plazo algunos de estos efectos pueden atenuarse, la transición sin medidas de acompañamiento provoca pérdidas sociales y personales considerables (Pantea et al., 2017).

Otro aspecto clave es el papel de la IA en la transformación de la innovación dentro de las empresas. Si bien puede agilizar procesos y aumentar la eficiencia, su implementación desigual puede reforzar disparidades ya existentes. En entornos donde la brecha de ingresos es alta, las empresas con más recursos capturan la mayor parte del beneficio tecnológico, mientras que aquellas con menor capacidad de inversión quedan rezagadas. Esto produce un círculo vicioso donde la innovación alimenta aún más la concentración económica y el poder del mercado (X. Li & Zhu, 2025).

3.3 Robótica avanzada

En las últimas décadas, el avance de robots impulsados por inteligencia artificial ha tenido un gran impacto en la sociedad, transformando profundamente las relaciones laborales, culturales, médicas y educativas. Sin embargo, esta transformación no ha sido unívocamente positiva, ya que también plantea importantes desafíos éticos, culturales y socioeconómicos.

En el mundo laboral, la automatización robótica está reemplazando muchas tareas repetitivas y peligrosas, mejorando la eficiencia operativa de las empresas. Sin embargo, este avance también ha generado preocupación, en especial la automatización de tareas rutinarias ha comenzado a reconfigurar el mercado laboral, afectando principalmente a trabajadores con menor cualificación. Aunque surgen nuevos roles vinculados al diseño, programación y mantenimiento de estos sistemas, no todos los sectores ni todas las regiones tienen acceso equitativo a estas oportunidades. La robótica asociada a la IA no solo automatiza procesos, sino que también redefine roles humanos, lo que exige una actualización constante de habilidades y, una adaptación de marcos normativos y de protección (Dumlao et al., 2025; Ribeiro et al., 2021).

Además del entorno laboral, los robots sociales están generando un impacto en la forma en la que los humanos se relacionan con las máquinas. La manera en que las personas perciben e interactúan con estos robots varía según factores culturales y contextuales. En algunas culturas, se tiende a proyectar emociones y atributos humanos en las máquinas, fenómeno conocido como antropomorfismo, lo que facilita la aceptación de robots como asistentes sociales o de acompañamiento. Sin embargo, este fenómeno también puede generar expectativas irreales sobre la empatía o la moralidad de las máquinas, y plantea interrogantes éticos sobre el vínculo emocional con entes no humanos (Roselli et al., 2025).

En el campo de la medicina, los robots quirúrgicos apoyados por IA se están convirtiendo en aliados fundamentales para mejorar la precisión, reduciendo los riesgos para los pacientes, los tiempos de recuperación y reducir la carga de trabajo en procedimientos complejos. Estos sistemas han demostrado su utilidad en situaciones donde los profesionales humanos enfrentan limitaciones físicas o mentales, como la fatiga por privación del sueño, contribuyendo así a una mayor seguridad en las intervenciones clínicas. Por otra parte, los robots de asistencia también están ayudando en la atención de personas mayores, pacientes crónicos o personas con movilidad reducida,

favoreciendo su autonomía y reduciendo la sobrecarga del personal sanitario. Una preocupación creciente sobre la pérdida de habilidades manuales por parte de personal médico, y la exclusión de hospitales con menos recursos para acceder a esta tecnología (Frieberg et al., 2025; B. S. Peters et al., 2018).

En el ámbito educativo y terapéutico, los robots de telepresencia están adoptando un rol cada vez más creciente como el apoyo emocional y académico, especialmente en contextos de enfermedad prolongada o aislamiento social. En experiencias recientes, se ha comprobado que estos dispositivos pueden facilitar la continuidad de la educación en los niños que no pueden asistir presencialmente a la escuela, reduciendo el aislamiento social y mejorando su bienestar emocional. Esta dimensión positiva de la robótica pone en evidencia su potencial para generar entornos más inclusivos y resilientes (Vegeberg et al., 2025).

Sin embargo, el desarrollo de robots autónomos también ha generado reacciones sociales ambiguas, especialmente cuando su presencia se percibe como una amenaza. La coexistencia entre humanos y robots puede desencadenar conflictos simbólicos, resiliencias culturales y tensiones laborales, particularmente cuando se teme que las máquinas desplacen a los trabajadores por completo, o que sus decisiones autónomas escapen del control humano. Este tipo de preocupaciones ha sido vinculado con el aumento de actitudes defensivas o incluso hostiles hacia robots en determinados contextos sociales (Roselli et al., 2025).

Además, el trabajo conjunto entre humanos y robots en entornos híbridos, como en hoteles, restaurantes, tiendas o bancos, está transformando la experiencia del cliente. Como se muestra en la Figura 6, el número de robots de servicios profesionales instalados anualmente a nivel global ha aumentado entre 2021 y 2023 con variaciones relevantes según el área de aplicación. Estos equipos mixtos pueden fomentar comportamientos de colaboración en la creación de valor, pero también pueden generar situaciones de frustración o desconfianza cuando los robots no logran responder de manera efectiva a las expectativas sociales o emocionales de los usuarios. La calidad de la interacción humano-robot no solo depende de la tecnología, sino también del diseño de experiencias significativas y adaptadas a las necesidades humanas (M. Li et al., 2025; Liu et al., 2025).

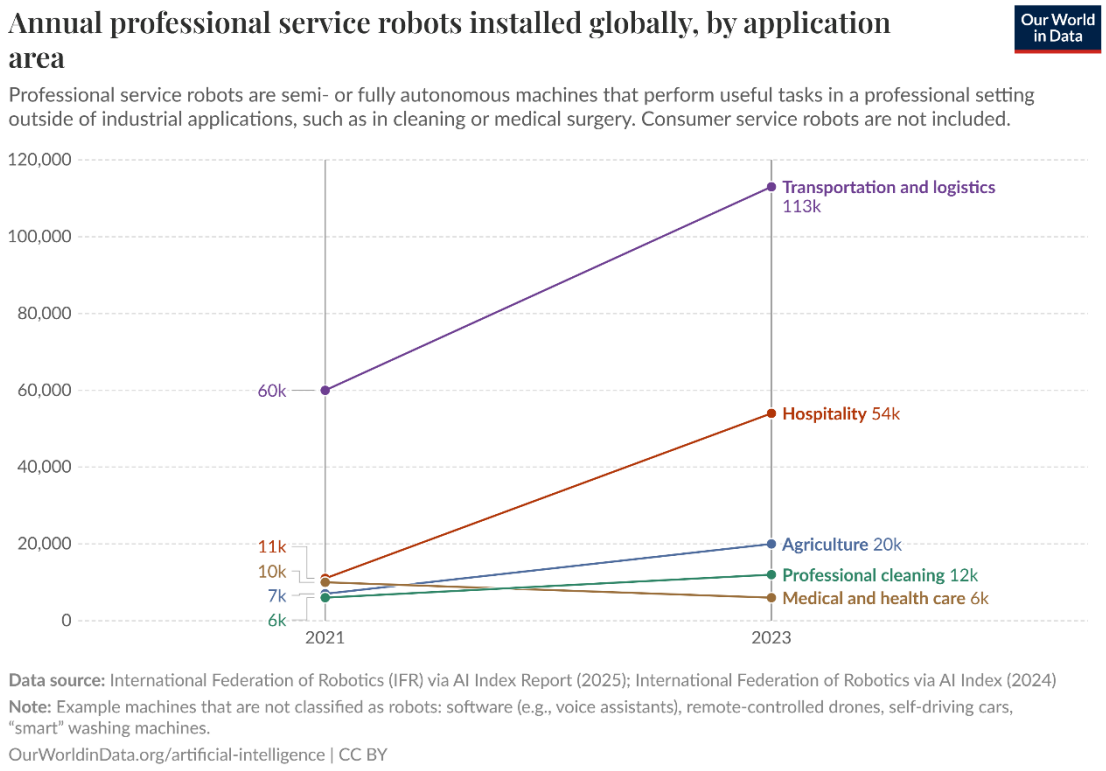

Figura 6. Robots de servicios profesionales instalados anualmente a nivel global, por área de aplicación (2021–2023)

Fuente: (Our World in Data, 2024)

El desarrollo de IA en robótica también plantea desafíos con la privacidad, la vigilancia y el control de datos personales. A medida que los robots se integran en espacios públicos y privados, recolectando y procesando informaciones, surgen preguntas urgentes sobre el uso ético de estos datos, la transparencia algorítmica y la protección de los derechos individuales. La presencia constante de sensores y sistemas inteligentes exige una gobernanza tecnológica que equilibre la innovación con la responsabilidad social (Ribeiro et al., 2021).

En términos más amplios, la expansión de los robots inteligentes ha generado reacciones sociales diversas. Algunos sectores los ven como una amenaza a los empleos y las relaciones humanas, mientras que otros valoran su potencial para aliviar cargas físicas y emocionales. Esta ambivalencia ha derivado en conflictos sociales, especialmente en regiones donde la llegada de robots ha sido percibida como invasiva o deshumanizante. La manera en que una sociedad reacciona ante estas tecnologías está influida por factores económicos, culturales y políticos, lo que requiere una atención interdisciplinaria para garantizar una implementación justa y responsable (Ribeiro et al., 2021).

3.4 Salud y bienestar

La inteligencia artificial ha revolucionado el campo de la salud, especialmente en imagenología médica, donde ha mejorado notablemente la capacidad diagnóstica y pronóstico de enfermedades complejas, como el cáncer. Las técnicas de aprendizaje automático y redes neuronales profundas se han empleado para analizar grandes volúmenes de datos radiológicos, permitiendo clasificar tumores según su modalidad de imagen y características histológicas con una precisión cada vez mayor. Este avance no solo agiliza el proceso diagnóstico, sino que también contribuye a una mejor estratificación del pronóstico, lo que favorece la personalización del tratamiento para cada paciente (Wu et al., 2021).

En radiología la IA ha demostrado su utilidad al facilitar la detección precoz y la caracterización de lesiones, reduciendo la carga de trabajo de los especialistas y mejorando la consistencia en la interpretación de imágenes médicas. La automatización de estas tareas permite una evaluación más rápida y detallada que resulta esencial en condiciones donde la rapidez en el diagnóstico es crucial como en los tumores malignos (Hosny et al., 2018). Además, los avances en técnicas de radiómica, que extraen características cuantitativas de las imágenes combinadas con aprendizaje profundo, ofrecen nuevas vías para estratificar el riesgo y pronósticos en cánceres, como el de pulmón, permitiendo diseñar tratamientos más ajustados a las características específicas de cada tumor (Cho et al., 2021).

El impacto de IA también se extiende a los programas de cribado de cáncer de mama, donde las herramientas automatizadas apoyan la identificación temprana de anomalías en mamografías. Estas tecnologías aumentan la sensibilidad del diagnóstico sin provocar un aumento significativo en los falsos positivos, lo que se traduce en mejores resultados clínicos y un uso más eficiente de los recursos médicos. El uso de la IA en este campo contribuye a mejorar la cobertura y eficacia de los programas preventivos, algo fundamental para reducir la mortalidad asociada a esta enfermedad (Álvarez-Benito et al., 2025; Díaz et al., 2021).

La IA está adquiriendo una posición más central dentro del sistema sanitario, con un crecimiento tanto en inversión como en desarrollo de aplicaciones clínicas. Esta tendencia refleja el reconocimiento del potencial transformador de la IA en la mejora de la eficiencia del sistema de salud, el desarrollo de tecnologías diagnósticas avanzadas y la personalización del tratamiento (Cicek & Bagci, 2025).

Un ámbito de especial relevancia es la monitorización remota de pacientes. Gracias a sensores inteligentes y modelos predictivos se pueden detectar signos de empeoramiento en enfermedades crónicas, oncológicas o incluso durante tratamientos de dolor persistente, lo que permite actuar antes de que la situación clínica se agrave (Patel et al., 2024). Estas herramientas son fundamentales en cuidados intensivos, donde sistemas basados en IA analizan señales fisiológicas en tiempo real para identificar condiciones críticas, como la sepsis (Mollura et al., 2021).

La inteligencia artificial ha comenzado a transformar la cirugía no solo a través de robots físicos, sino mediante sistemas de aprendizaje profundo que asisten en la planificación y ejecución de operaciones. Estas herramientas permiten identificar estructuras anatómicas en tiempo real, detectar riesgos y ofrecer apoyo en decisiones intraoperatoria mejorando así la precisión en la seguridad del procedimiento (Morris et al., 2023).

Además, la IA facilita la predicción de complicaciones, la personalización de planes quirúrgicos y la evaluación objetiva del rendimiento del cirujano, lo que puede optimizar tanto la formación médica como los resultados clínicos. También ha demostrado reducir tiempos quirúrgicos y mejorar la eficiencia operativa, especialmente en cirugías complejas como las oncológicas. Sin embargo, su implementación aún enfrenta desafíos, como la interoperabilidad tecnológica, la necesidad de datos quirúrgicos de calidad y cuestiones éticas sobre la delegación de decisiones críticas (Fuentes et al., 2025).

La expansión de la telemedicina, impulsada por tecnologías basadas en inteligencia artificial, ha redefinido la relación entre pacientes y sistemas de salud. La incorporación de algoritmos inteligentes en plataformas de atención remota ha permitido mejorar el diagnóstico temprano, la continuidad del cuidado y la toma de decisiones clínicas en tiempo real. Estas tecnologías han demostrado un especial potencial durante la crisis sanitaria como la pandemia de COVID-19, en las que la atención presencial resultó inviable o peligrosa. No obstante, el despliegue global de la telemedicina también ha revelado desigualdades en el acceso de la infraestructura digital y conectividad, lo cual limita su impacto beneficioso en regiones con escasos recursos tecnológicos. Además, la falta de marcos regulatorios sólidos y la escasa estandarización internacional generan incertidumbre sobre la calidad, la equidad y la seguridad de estos servicios, lo que refuerza la necesidad de políticas públicas inclusivas y colaborativas que garanticen su implementación ética y sostenible (Su et al., 2024).

Es necesario garantizar que los algoritmos sean rigurosamente validados y que su funcionamiento sea transparente para evitar sesgos que puedan afectar la equidad en el acceso y la calidad de la atención médica. Asimismo, la protección de datos sensibles y la ética en la implementación de estas tecnologías son aspectos que requieren atención prioritaria para preservar la confianza de los pacientes y profesionales sanitarios (Basáez & Mora, 2022; Koh et al., 2022).

3.5 Aplicaciones en salud mental

El uso de la inteligencia artificial en salud mental representa uno de los desarrollos tecnológicos más prometedores, y también más sensible, del siglo XXI. Lejos de limitarse al diagnóstico automatizado o a la digitalización de expedientes clínicos, la IA está emergiendo como una herramienta transformadora con un impacto profundo en la forma que comprendemos, abordamos y gestionamos el bienestar psicológico de las personas y las comunidades. Este impacto se manifiesta en múltiples dimensiones desde la atención personalizada pasando por el acceso a los servicios hasta efectos culturales y éticos en las relaciones terapéuticas y la concepción misma del sufrimiento humano.

Uno de los impactos sociales más evidentes es la democratización del acceso a servicios psicológicos. En contextos dónde hay escasez de profesionales de salud mental o dónde persiste un estigma fuerte hacia la psicoterapia tradicional, las aplicaciones y las plataformas de ellas están facilitando el acceso a intervenciones básicas. Algunas herramientas como chatbots terapéuticos o aplicaciones de autoevaluación emocional permiten a miles de personas, especialmente jóvenes y usuarios digitales, expresar sus preocupaciones y recibir estrategias de afrontamiento sin necesidad de contacto presencial (D'Alfonso, 2020).

Esta accesibilidad inmediata y continua, ha tenido un efecto directo en la reducción de barreras estructurales como los altos costes, las listas de espera o las limitaciones geográficas. La posibilidad de recibir orientación emocional a cualquier hora y en cualquier lugar ha transformado la forma en que muchas personas se relacionan con su salud mental. Además, estas plataformas han comenzado a naturalizar el uso de tecnologías como parte de las rutinas de autocuidado emocional, lo que contribuye a una normalización social del diálogo sobre bienestar psicológico. Casos como Woeboy o Wysa han demostrado resultados positivos en la reducción de síntomas leves a moderados de ansiedad y depresión, especialmente en jóvenes o personas que aún no habían buscado ayuda profesional (Horn & Weisz, 2020).

Otro ámbito en el que la IA ha tenido un impacto social relevante es la atención de poblaciones envejecidas, especialmente aquellas en situación de fragilidad o aislamiento. Actualmente se utilizan algoritmos que monitorean patrones de comportamiento, expresiones lingüísticas o cambios sutiles en el habla para identificar signos tempranos de deterioro cognitivo, depresión o ansiedad en personas mayores. Estos sistemas están siendo incorporados en programas de atención geriátrica y permiten intervenciones más rápidas y personalizadas, ayudando a prevenir situaciones de dependencia o crisis emocionales en etapas avanzadas de la vida (Wei et al., 2023).

La práctica clínica psicológica también está siendo alterada por la integración de la IA herramientas que analizan automáticamente textos de las sesiones y reconocen patrones en el lenguaje emocional o generan indicadores sobre el progreso del tratamiento, están siendo utilizadas por terapeutas como apoyo para la toma de decisiones. Esto ha permitido una gestión más eficiente del tiempo y mejor seguimiento de los pacientes y una evaluación más sistematizada de los tratamientos. Sin embargo, ha generado debate sobre la posibilidad de examinación del proceso terapéutico y la necesidad de preservar la dimensión relacional en la atención psicológica (J. Zhao et al., 2022).

Además, el uso extendido de tecnologías inteligentes está modificando la percepción social de lo que significa recibir ayuda psicológica. Muchas personas, especialmente jóvenes adultos, han comenzado a entender el acompañamiento emocional como algo que puede darse también a través de una interfaz digital sin que eso implique necesariamente menor calidad. Esta transformación cultural está configurando las expectativas sociales y afectivas en torno al cuidado mental, al mismo tiempo que, plantea preguntas sobre la privacidad, la confidencialidad de los datos y la ética de las intervenciones automatizadas (Graham et al., 2019).

En entornos clínicos especializados, también se está observando como la IA mejora la detección de trastornos mentales complejos, gracias a su capacidad para identificar correlaciones entre síntomas, historial médico, conducta verbal y los datos biométricos. Esto ha permitido detectar casos de depresión, ansiedad o trastornos del estado de ánimo que anteriormente pasaban desapercibidos, y ha facilitado la atención temprana, especialmente en servicios saturados (de la C. Hernández Lugo et al., 2025). A través de algoritmos de aprendizaje automático entrenados en grandes conjuntos de datos, es posible identificar signos de depresión, ansiedad, estrés postraumático o incluso riesgo de suicidio analizando variables como el tono de voz, los patrones de escritura, las

expresiones faciales, el ritmo del sueño o la actividad en redes sociales (D'Alfonso, 2020; Graham et al., 2019).

A pesar de sus beneficios el uso extendido de la inteligencia artificial, en el ámbito de la salud mental también está teniendo consecuencias sociales adversas. Una de las principales preocupaciones actuales es que el apoyo psicológico mediado por sistemas automatizados está reduciendo la necesidad de interacción humana directa. Muchas personas, especialmente jóvenes, optan por recurrir a chatbots o aplicaciones de IA para desahogarse o buscar orientación emocional, lo que disminuye su exposición a la empatía humana y al diálogo interpersonal profundo. Este cambio puede fomentar el retraimiento social, el debilitamiento de las habilidades cognitivas y la preferencia por soluciones tecnológicas, en lugar de vínculos afectivos reales. Además, al recibir respuestas prefabricadas asistidas por algoritmos se limita al desarrollo de la creatividad emocional y la capacidad de afrontar situaciones complejas de forma espontánea. Este fenómeno sumado al aumento del tiempo en entornos digitales se relaciona con mayores índices de soledad, desconexión social y pérdida de sentido de comunidad (de la C. Hernández Lugo et al., 2025; Graham et al., 2019).

3.6 Movilidad y transporte

En la actualidad, la inteligencia artificial ya está generando efectos sociales tangibles en el ámbito de la movilidad. Uno de los más visibles es su integración en sistemas inteligentes de transporte urbano, como la regulación activa del tráfico, la predicción de congestiones y la optimización de rutas. Estas tecnologías están modificando la experiencia cotidiana de movilidad en muchas ciudades al reducir los tiempos de desplazamiento y mejorar la eficiencia general del sistema vial (Mauro et al., 2025).

Sin embargo, estos avances no están exentos de consecuencias sociales relevantes. A medida que se incorporan algoritmos de recomendación en la planificación urbana y de transporte, surgen efectos de retroalimentación que pueden reforzar patrones de movilidad preexistente y reproducir desigualdades espaciales, favoreciendo las zonas con mayor densidad de datos o mejor infraestructura digital. Todo esto plantea riesgos de exclusión para comunidades menos conectadas o infrarrepresentadas en los sistemas de IA (Bang et al., 2024; Mauro et al., 2025).

Otro efecto observable es la transformación de la relación usuario-sistema. Algunas tecnologías como los sistemas de navegación inteligentes o los vehículos con asistencia parcial están desplazando parte del control, desde el conductor hacia el sistema automatizado, alterando el rol activo del usuario y generando nuevos desafíos sobre la

confianza, la atención y la responsabilidad en la conducción. El diseño sociotécnico de vehículos autónomos requiere un enfoque multidimensional que integre aspectos humanos como confianza y transparencia, tecnológicos como robustez, y regulatorios para evitar fallos o rechazos por parte de usuarios con baja familiaridad tecnológica (Gilbert et al., 2022).

Además, comienzan a evidenciarse impactos laborales derivados de la automatización parcial de tareas en sectores como el transporte de mercancías y pasajeros, lo que puede provocar la restructuración funcional del trabajo. Aunque no se ha producido aún una sustitución masiva de empleos, sí se observan procesos de reconversión y desplazamientos en servicios logísticos y de transporte urbano que requieren nuevas competencias digitales y adaptabilidad (Bang et al., 2024; Bucchiarone et al., 2020).

3.7 Educación

La incorporación de la inteligencia artificial en la educación ha tenido importantes repercusiones sociales, redefiniendo los modelos de acceso al conocimiento, la relación entre estudiantes y la tecnología, y las expectativas sobre el rol de la enseñanza. Uno de los principales impactos positivos ha sido la ampliación del acceso a recursos educativos personalizados. Gracias a algoritmos adaptativos y plataformas inteligentes los estudiantes pueden recibir apoyo ajustado a su ritmo y estilo de aprendizaje, lo que contribuye a una mayor equidad en el acceso a la educación (Fütterer et al., 2023).

Además, la IA ha potenciado la inclusión digital al facilitar la traducción automática, la generación de contenidos accesibles y el soporte a estudiantes con necesidades especiales, contribuyendo a una mayor participación social en entornos educativos diversos. También ha favorecido el aprendizaje autónomo y fuera del aula permitiendo que el conocimiento se distribuya de manera descentralizada y flexible, lo que ha reforzado la noción de aprendizaje a lo largo de la vida (Jiang et al., 2024)

Desde una perspectiva social más amplia la presencia de la IA en la educación ha impulsado un debate sobre la alfabetización digital, la ética tecnológica y en la necesidad de preparar los ciudadanos no solo como usuarios, sino como agentes críticos capaces de comprender y cuestionar los algoritmos que influyen en procesos de aprendizaje. Esto ha generado un cambio cultural significativo, en el que las habilidades digitales se perciben ya no solo como complementarias, sino como competencias clave para la participación plena en la sociedad del conocimiento (García-Peñalvo, 2023).

No obstante, también existen efectos negativos el acceso desigual a tecnologías basadas en la IA puede acentuar las brechas existentes entre estudiantes

particularmente en regiones con escasa infraestructura digital, esta desigualdad tecnológica puede traducirse en nuevas formas de exclusión educativa afectando especialmente a grupos vulnerables (Fütterer et al., 2023). Por otro lado, la posible automatización de ciertas tareas académicas como la evaluación o la generación de contenidos ha despertado inquietudes sobre la posible diseminación de los procesos educativos y la pérdida de habilidades críticas y creativas en los estudiantes (Jiang et al., 2024).

Asimismo, la creciente presencia de los sistemas de IA en entornos educativos ha generado tensiones respecto al papel del profesorado, si bien estas tecnologías pueden actuar como herramientas de apoyo, también existe el riesgo de que se diluya la función pedagógica del docente, sustituyéndola por soluciones algorítmicas sin el componente humano indispensable para el aprendizaje significativo (García-Peñalvo, 2023).

3.8 Entretenimiento

Uno de los ámbitos donde la inteligencia artificial está generando transformaciones sociales profundas y, al mismo tiempo, está suscitando preocupaciones éticas es en el entretenimiento digital. Desde plataformas de música y vídeos hasta redes sociales y videojuegos, la IA desempeña un papel central en la creación, distribución y personalización de contenidos, alterando tanto las dinámicas de producción como de consumo cultural. Los sistemas de recomendación basados en IA, diseñados para maximizar el tiempo de atención del usuario, funcionan mediante el análisis de grandes volúmenes de datos y patrones de comportamiento. Aunque logran ofrecer experiencias altamente personalizadas, estos algoritmos tienden a reforzar los hábitos persistentes del usuario, disminuyendo la exposición a contenidos diversos y alimentando fenómenos como la formación de cámaras de eco, burbujas de filtro y radicalización algorítmica (Karnouskos, 2020; Rodilosso, 2024).

Este fenómeno no solo es problema de diseño técnico, sino también plantea interrogantes sobre el pluralismo cultural y la responsabilidad de las plataformas digitales. Si los usuarios solo acceden al contenido que confirma sus creencias y gustos se reduce su capacidad crítica y se debilita la libre elección. Esto es particularmente problemático en contextos donde el entretenimiento y la información convergen, como en las redes sociales o los formatos híbridos de contenido, dónde la distinción entre lo lúdico y lo político puede volverse difuso. Además, la priorización algorítmica de contenidos optimizados para la atención puede favorecer lo emocionalmente extremo,

lo sensacionalista o lo polémico, desplazando contenidos más complejos, diversos o culturalmente relevantes (Rodilosso, 2024; Z. Zhao et al., 2024).

Paralelamente, la aparición y proliferación de tecnologías de generación audiovisual como los deepfakes ha provocado un cambio radical en la producción de imágenes y narrativa. En el mundo del cine y la televisión, los deepfakes se han utilizado para rejuvenecer actores, recrear personajes fallecidos o modificar escenas de forma eficiente y económica. Ejemplos como la película de *The Irishman*, donde se empleó esta tecnología para representar a los actores en distintas edades, han demostrado su potencial creativo (Debroy et al., 2024). Sin embargo, esta misma capacidad de manipulación visual ha generado alarmas en torno a la autenticidad y la integridad de los contenidos audiovisuales. La posibilidad de crear imágenes o vídeos hiperrealistas de personas diciendo o haciendo cosas que nunca ocurrieron plantea serias amenazas para la privacidad, la reputación y la percepción pública (Broinowski & Martin, 2024; Rodilosso, 2024).

Desde el punto de vista psicológico y social, los contenidos falsos generados por IA afectan no solo la percepción individual, sino el comportamiento colectivo. La dificultad para distinguir lo real de lo ficticio puede generar desconfianza generalizada, ansiedad o confusión, especialmente cuando estas tecnologías se utilizan para desinformar, manipular emociones o socavar la credibilidad de personas e instituciones (Lundberg & Mozelius, 2024). Además, la constante exposición a estos contenidos puede dañar los estándares habituales y aumentar la sensación de no alineación o desconexión con la realidad. Estudios recientes alertan de este tipo de contenidos pueden afectar la salud mental, promoviendo emociones negativas como paranoia, frustración o sensación de vulnerabilidad, especialmente cuando se vulnera la identidad o la imagen de individuos sin su consentimiento (Debroy et al., 2024).

Frente a estos desafíos, algunos investigadores y diseñadores proponen enfoques centrados en la diversidad y en la ética desde la fase de diseño de los sistemas. En el caso de las plataformas musicales, por ejemplo, ya se están desarrollando algoritmos de recomendación que incorporan criterios de diversidad cultural, pluralidad de estilos o representación de minorías, con el objetivo de contrarrestar los efectos homogeneizantes del filtrado colaborativo (Porcaro et al., 2021). De manera complementaria, se están desarrollando iniciativas regulatorias que promuevan la transparencia algorítmica, el consentimiento informado para el usuario de datos e imágenes, y mecanismos de verificación para evitar propagación de contenidos manipulados. Estas normativas son aún emergentes, pero representan un paso

necesario para proteger los derechos de los usuarios en un entorno mediático cada vez más influido por la inteligencia artificial (Broinowski & Martin, 2024).

4 Retos, peligros y desafíos de la Inteligencia Artificial

La creciente integración de la inteligencia artificial en distintos sectores de la sociedad no solo ha generado oportunidades, sino una serie de riesgos y desafíos que requieren atención. Desde problemas éticos hasta posibles amenazas a los derechos fundamentales, la implementación de sistemas de IA implica asumir responsabilidades técnicas, legales y sociales.

Antes de profundizar en los riesgos, es crucial definir los conceptos clave que guían el análisis realizado para entender el alcance de los riesgos.

La confianza, en este contexto, puede entenderse como la expectativa que un sistema automatizado actúa de forma coherente, predecible y en línea con los intereses del usuario. No obstante, algunos autores advierten que esta confianza no es equivalente a la que se depositan los seres humanos, ya que las máquinas no poseen intenciones ni conciencia moral (Søgaard, 2023).

La confiabilidad hace referencia a la capacidad técnica de un sistema para desempeñar su función sin fallos durante un periodo determinado. En los sistemas de IA implica que el sistema debe responder correctamente bajo condiciones específicas y con un rendimiento estable (Budnik, 2025). A menudo, se emplea también el término fiabilidad, que además de la dimensión técnica, incorpora la percepción del usuario respecto al comportamiento del sistema, incluso si un sistema es confiable en términos operativos y el usuario no lo percibe como tal, su fiabilidad práctica puede verse afectada (Simion & Kelp, 2023). En esta línea, algunos autores sostienen que no es necesario diseñar sistemas moralmente confiables como si fueran humanos, sino que basta con que comuniquen de forma clara y comprensible su nivel de confiabilidad, lo cual permitiría a los usuarios ajustar racionalmente su grado de confianza (Dorsch & Deroy, 2024)

El concepto de vulnerabilidad hace referencia a la exposición de un sistema a posibles daños, errores o ataques. Esta vulnerabilidad no solo es técnica, sino también social y ética, ya que puede derivar en efectos no deseados como discriminación, falta de transparencia o pérdida de control humano (Reinhardt, 2022).

En este apartado se abordará tanto la futura legislación, como los riesgos actuales y los posibles riesgos futuros asociados a esta tecnología.

4.1 Riesgos actuales

La rápida expansión de la inteligencia artificial ha traído consigo una serie de riesgos que requieren atención urgente. Estos desafíos afectan tanto la privacidad como la igualdad, la autonomía humana o la rendición de cuentas. Además, la falta de regulación clara y una creciente dependencia tecnológica agravan un panorama complejo que requiere una reflexión crítica y una gestión responsable.

4.1.1 Problemas de privacidad y seguridad

Uno de los riesgos más significativos asociados al desarrollo de la inteligencia artificial es la falta de privacidad y la seguridad de los datos personales. A medida que los sistemas de IA se integran en procesos sociales, políticos y económicos, crece la preocupación sobre el acceso, uso y almacenamiento de información sensible por la sociedad.

Los sistemas de IA para funcionar de manera eficiente requieren grandes volúmenes de datos con los que entrenar a los algoritmos. Esto ha llevado a prácticas cada vez más intrusivas de recopilación de información, muchas veces sin el consentimiento explícito de los usuarios o sin garantizar que los datos serán tratados de forma segura y ética. En contextos democráticos este fenómeno representa un riesgo doble: por un lado, amenaza los derechos individuales de la privacidad, y por otro, puede ser utilizado como herramienta para influir o manipular procesos electorales, socavando la integridad democrática (Shams, 2025).

La regulación legal en materia de protección de datos aún no ha logrado adaptarse completamente al ritmo de avance de la IA. Aunque existen marcos normativos como el Reglamento General de Protección de Datos (RGPD) en Europa, estos todavía presentan lagunas cuando se aplican a tecnologías emergentes como el aprendizaje automático, especialmente lo que representa a la garantía del anonimato de los datos, el consentimiento informado y la explicación de decisiones automatizadas (Kathuria et al., 2024). Estas debilidades normativas permiten que entidades con capacidad tecnológica significativa puedan explotar vacíos legales para recolectar, analizar y comercializar datos personales sin una supervisión adecuada.

Casos recientes que ilustran estos riesgos, como la empresa Luka Inc, que desarrolla el chatbot IA Replika, fue multada con 5.6 millones de dólares por las autoridades italianas por procesar datos personales sin una base legal válida y por no implementar controles efectivos de verificación de edad que derivó en una

exposición de contenido inapropiado a menores de edad. Otro caso sancionado es el de Clearview AI que fue multada con 30.5 millones de euros por recopilar imágenes faciales sin consentimiento, infringiendo directamente el RGPD (Autoriteit Persoonsgegevens, 2024). Estas prácticas muestran como el uso indebido de datos puede vulnerar derechos fundamentales a gran escala.

Por otro lado, los riesgos de seguridad son una preocupación central. Microsoft, por ejemplo, filtró accidentalmente 38 TeraBytes de datos privados al compartir un conjunto de datos de entrenamiento de IA que incluía contraseñas, claves privadas y mensajes internos, evidenciando lo frágil que puede ser la gestión de información confidencial en entornos de IA (Culafi, 2023). A esto se suma que el 84% de las herramientas de IA analizadas en un estudio reciente habían sufrido brechas de seguridad y el 51% había enfrentado robos de sus credenciales, generalmente por el uso sin supervisión y configuraciones deficientes de estas herramientas (Graterol, 2025).

Además, las diferencias legislativas entre jurisdicciones complican aún más la protección de datos en un entorno globalizado. Mientras que algunas regiones promueven enfoques estrictos y centrados en los derechos del usuario, otras adoptan políticas más permisivas que priorizan la innovación tecnológica sobre la privacidad individual, lo que genera una descompensación legal que pueden ser aprovechadas por multinacionales (Kathuria et al., 2024).

4.1.2 Sesgo algorítmico

Uno de los desafíos éticos más relevantes en el desarrollo de sistemas de inteligencia artificial es el riesgo a que se reproduzcan y agraven las desigualdades sociales preexistentes. Este fenómeno, conocido por sesgo algorítmico, ocurre cuando los modelos aprenden de datos que contienen prejuicios históricos o reflejan dinámicas estructuralmente discriminatorias, perpetuando así resultados injustos. Estos efectos suelen afectar con mayor intensidad a colectivos históricamente marginados, como mujeres, personas racializadas, personas migrantes e individuos con bajos ingresos o en situación de pobreza.

Este problema no se limita únicamente a los datos con los que se entrena al algoritmo. También intervienen decisiones técnicas como la selección de variables y la arquitectura del modelo, así como los aspectos sociales y la composición homogénea de los equipos de desarrollo. Estos dificultan la detección y corrección de los sesgos, especialmente cuando los modelos se

aplican a ámbitos sensibles como la justicia penal, los procesos de contratación o la concesión de créditos (Ryan, 2020). Esto puede llevar, por ejemplo, a que personas racializadas reciban evaluaciones de riesgo más altas y con posibilidad de penas mayores en procesos judiciales automatizados o que las mujeres sean sistemáticamente excluidas de procesos de selección en sectores dominados por hombres.

Además de sus implicaciones éticas, el sesgo algorítmico conlleva consecuencias epistemológicas relevantes. Un sistema que aprende a partir de datos distorsionados no solo reproduce decisiones injustas, sino que también genera un conocimiento defectuoso. La calidad de los resultados producidos por la inteligencia artificial depende directamente de la fiabilidad de los datos que utiliza, cuando estos datos están sesgados, el conocimiento resultante se ve comprometido, lo que afecta a su aplicabilidad en campos como la ciencia, la salud o las políticas públicas (Russo et al., 2024).

Este panorama exige reflexionar el papel que desempeñan los seres humanos en el desarrollo y supervisión de la IA. A medida que las tecnologías se vuelven más potentes, también aumenta la necesidad de asumir mayores niveles de responsabilidad sobre su funcionamiento y sus efectos sociales. La complejidad técnica no puede ser una excusa para renunciar al juicio ético, sino un motivo para reforzarlo (De Cremer & Kasparov, 2021). Esto implica fomentar una mayor diversidad en los equipos de desarrollo de IA, aplicar auditorías externas y establecer marcos normativos que aseguren la equidad en el diseño y aplicación de algoritmos.

4.1.3 Opacidad de la IA

La opacidad en los sistemas de inteligencia artificial constituye un riesgo ético clave, comúnmente conocido como “el problema de caja negra”, ya que limita la compresión y la evaluación de decisiones automatizadas. Esta opacidad puede diferenciarse en dos formas principales: por un lado, como una limitación epistémica que impide entender cómo se alcanzan algunas conclusiones; por otro lado, metodologicamente como una falta de trazabilidad y la dificultad en el entrenamiento y configuración del sistema (Boge, 2022). Esta doble dimensión compromete tanto la transparencia como la capacidad de control humano, que plantea ya no solo riesgos técnicos, sino morales y sociales, ya que limita la posibilidad de evaluar críticamente las decisiones tomadas por los sistemas.

Algunos autores defienden que la opacidad no implica necesariamente una falta de legitimidad ética, los sistemas opacos pueden ser éticamente aceptables si ofrecen resultados fiables y están sometidos a mecanismos de supervisión análogos y adecuados a los criterios que aplicamos al confiar en expertos humanos (Ryan, 2020). No obstante, esta comparación es controvertida, ya que los sistemas de IA carecen de intención, de juicio moral y el contexto que caracteriza las decisiones humanas.

En sectores como la medicina, el riesgo se acentúa y es un problema que cobra especial relevancia. La creciente delegación de decisiones clínicas en sistemas de IA puede reducir la capacidad crítica de los profesionales sanitarios y generar una dependencia que disminuya su conocimiento práctico, y con ello su autonomía y competencia, esto es definido como una pérdida epistémica significativa (Schmidt et al., 2025).

La opacidad no solo amenaza la transparencia técnica, sino también la rendición de cuentas, la autonomía profesional y la confianza pública, por lo que debe ser tratada como un problema ético estructural que requiere respuestas institucionales y tecnológicas.

4.1.4 Manipulación de información y desinformación

Ante la proliferación del uso de la inteligencia artificial en todos los ámbitos ha transformado radicalmente la forma en la que se produce distribuye y consume la información, facilitando tanto la creación como la detección de contenidos falsos. Esta dualidad plantea desafíos significativos para la integridad informativa y para la confianza pública.

La tecnología de deepfakes, basada en inteligencia artificial generativa, ha emergido como uno de los desarrollos más disruptivos en el ámbito digital actual. Esta técnica permite crear imágenes, audios y vídeos extremadamente realistas que representan a personas diciendo o haciendo cosas que nunca ocurrieron en la realidad. Si bien tiene aplicaciones positivas, como su uso en el cine, la accesibilidad o el entretenimiento educativo, los riesgos superan con frecuencia sus beneficios. Uno de los peligros más destacados es su capacidad para difundir desinformación, deteriorar la confianza en los medios y fomentar la suplantación de identidad en contextos personales, comerciales o legales. Además, se han documentado casos de deepfakes utilizados para el acoso, la extorsión y la creación de pornografía no consentida, afectando a figuras públicas y, en especial, a mujeres. Este avance tecnológico en este ámbito

supera la capacidad actual de las herramientas de detección, lo que genera un entorno de incertidumbre donde el contenido audiovisual puede perder su valor y su fiabilidad. Frente a este panorama, resulta fundamental fomentar la conciencia crítica de los usuarios, promover la educación digital desde edades tempranas y desarrollar sistemas más robustos de verificación para enfrentarse a estos desafíos con responsabilidad (Misirlis & Munawar, 2023).

Además, las redes sociales actúan como catalizadores en la propagación de desinformación, especialmente en temas de salud pública. Durante la pandemia de COVID-19, se observó como información errónea sobre tratamientos y vacunas se difundía rápidamente, afectando negativamente a las decisiones de salud de la población. Esta difusión de información masiva de cualquier tipo ha sido identificada como una amenaza significativa para la salud pública, exacerbada por la velocidad y el alcance de las plataformas digitales (Denniss & Lindberg, 2025).

La manipulación de información mediante IA también ha transformado las campañas políticas. El uso de Big Data y algoritmos permite a los partidos políticos realizar una microsegmentación de votantes, adaptando mensajes específicos para influir en sus decisiones. Esta práctica, conocida como microtargeting político, plantea preocupaciones éticas sobre la privacidad y la manipulación de los votantes (Tomić et al., 2023).

Además, la exposición repetida a información falsa, incluso cuando es identificada como tal, puede reforzar creencias erróneas debido al efecto de verdad ilusoria. Este fenómeno se ve amplificado por el uso de redes sociales como fuente de información de noticias, donde la desinformación se propaga rápidamente y con gran alcance (Ahmed et al., 2024).

Recientemente, se ha destacado que los deepfakes representan una nueva frontera de desinformación política. Esta tecnología sofisticada como la IA difumina cada vez más la línea entre la realidad y la ficción, siendo utilizada para manipular la opinión pública y erosionar la confianza en los medios de comunicación. Los deepfakes ya han sido empleados para atacar a figuras públicas, crear narrativas divisivas y comprometer la seguridad nacional, lo que genera preocupaciones sobre su potencia para socavar los procesos democráticos y fomentar la discordia social (Vaccari & Chadwick, 2020)

Frente a estos desafíos, la IA también ofrece soluciones prometedoras. Se han desarrollado sistemas que utilizan procesamiento de lenguaje natural y algoritmos avanzados para detectar y mitigar la difusión de noticias falsas. Sin

embargo, estos enfoques enfrentan obstáculos como la necesidad de supervisión humana continua, la constante actualización de los algoritmos y la adaptación de nuevas tácticas de desinformación. La colaboración entre los sectores y la integración de enfoques éticos, son esenciales para mejorar la eficacia de estas herramientas (Saeidnia et al., 2025).

4.1.5 Automatización cada vez más invasiva

La automatización impulsada por la inteligencia artificial está transformando rápidamente el panorama laboral, sustituyendo tareas humanas en sectores como atención al cliente, la programación y la administración. A pesar de sus beneficios evidentes, como la mejora de eficiencia y la reducción de costes, esta tendencia conlleva riesgos estructurales, económicos y sociales que requieren una evaluación critica.

Uno de los principales riesgos es el impacto negativo sobre el empleo. La automatización tiende a sustituir las tareas más repetitivas y previsibles que antes realizaban personas. Aunque, en algunos casos la IA, puede complementar el trabajo humano, aumentando la productividad, en muchos otros aspectos provoca la eliminación de puestos de trabajo esto genera preocupación sobre el futuro laboral de millones de personas especialmente en sectores donde la sustitución por máquinas es más factible (Wang & Lu, 2025).

Además, la problemática del desplazamiento laboral es un tema que ha ganado relevancia en los últimos años. No solo se trata de perder empleos, sino también de la dificultad que tienen muchos trabajadores para adaptarse a los nuevos requerimientos. Un análisis bibliográfico de las publicaciones científicas sobre automatización y empleo revela que el interés académico y social en el desplazamiento laboral ha crecido de forma constante en las últimas décadas, evidenciando la relevancia y complejidad del problema. Este estudio también destaca que, aunque se crean nuevos empleos relacionados con la tecnología, muchos trabajadores tienen dificultades para adaptarse debido a la falta de formación adecuada, lo que puede aumentar las desigualdades sociales y económicas. La automatización exige nuevas habilidades y competencias técnicas, pero la formación y la educación no siempre están al nivel necesario para cubrir esta demanda, lo que puede perpetuar dichas desigualdades (Subaveerapandiyan & Shimray, 2024).

Otro riesgo importante está relacionado con la concentración de beneficios. Las empresas tecnológicas y los profesionales con alta cualificación suelen ser los

principales beneficiarios de la automatización, mientras que los trabajadores menos especializados quedan en una situación de desventaja. Esto puede aumentar la brecha económica y social, dificultando la igualdad de oportunidades y generando tensiones sociales. Además, la falta de transparencia en los algoritmos y las decisiones automatizadas pueden afectar a la justicia y a la ética en el trabajo (Polak, 2021).

Por tanto, aunque la automatización ofrece ventajas claras en productividad y competitividad, es esencial implementar políticas públicas que mitiguen esos efectos negativos. Esto incluye promover la formación continua, establecer regulaciones claras y supervisar éticamente los sistemas automatizados para asegurar una transición justa e inclusiva hacia la digitalización.

4.1.6 Dependencia y deterioro de las capacidades humanas

A medida que la inteligencia artificial se integra cada vez más en la vida cotidiana, surgen riesgos que afectan negativamente al desarrollo humano, tanto a nivel individual como colectivo. Entre ellos, destaca la creciente dependencia de la tecnología y en este caso de la IA, la cual no solo nos facilita algunos ámbitos de la vida diaria, sino que también puede deteriorar las habilidades cognitivas, sociales y emocionales fundamentales del ser humano.

Uno de los peligros más evidentes es el uso excesivo de sistemas de IA en la comunicación diaria, como los chatbots, Aunque estos sistemas permiten mantener conversaciones fluidas, su uso frecuente está alterando la forma en la que las personas se comunican. Se ha observado que este tipo de interacción puede debilitar la percepción de la autenticidad de las conversaciones, generar desconfianza interpersonal que erosiona a la calidad del vínculo humano en la sociedad (Hohenstein et al., 2023). Este empobrecimiento de las relaciones sociales representa una amenaza para el desarrollo de habilidades sociales complejas y la posibilidad de sustituir vínculos sociales por interacciones artificiales, más funcionales que significativas.

Por otro lado, los sistemas de IA pueden influir directamente en la forma en que los individuos perciben la realidad. Estas herramientas tienden a confirmar creencias previas, reforzando sesgos existentes y limitando la capacidad crítica. El resultado es una percepción parcial y distorsionada del entorno social y emocional, lo que compromete al juicio personal y a la interpretación de situaciones cotidianas (Glickman & Sharot, 2024).

En el ámbito educativo, los efectos no son menos preocupantes. Si bien la IA puede facilitar el aprendizaje, su uso abusivo puede reducir la motivación para pensar por uno mismo. Al delegar tareas complejas a sistemas automáticos, los estudiantes corren el riesgo de perder habilidades importantes, como el análisis crítico, la resolución de problemas o la toma de decisiones autónoma. Esta dependencia excesiva compromete el desarrollo de la autonomía intelectual, especialmente en las etapas de formación (Jose et al., 2025; Zhai et al., 2024).

Desde el punto de vista psicológico, el uso intensivo de tecnologías basadas en IA se ha vinculado con niveles más altos de ansiedad, sensación de aislamiento y disminución del bienestar emocional, en parte por la sustitución de la interacción humana por vínculos funcionales con máquinas (Huang et al., 2024). Estas consecuencias afectan especialmente a los jóvenes que están más expuestos a estas tecnologías en su vida cotidiana.

El avance la IA no solo modifica el modo en el que se trabaja o se aprende, sino que también puede producir un deterioro profundo de las capacidades humanas, por lo que es necesario gestionar esta tecnología de forma crítica y responsable.

4.1.7 Implicaciones legales y normativas

La falta de un marco normativo claro y coherente sobre el uso de la inteligencia artificial constituye otro riesgo fundamental.

El desarrollo tan rápido de la inteligencia artificial ha superado ampliamente la capacidad de las normativas vigentes para regular su uso de forma adecuada. Esta falta de regulación, especialmente en lo relativo a la asignación de responsabilidades, la protección de derechos fundamentales y la legitimidad de las decisiones automatizadas (U. Peters, 2022). A diferencia de los humanos, los sistemas de IA no operan bajo marcos normativos compartidos, ni justifican sus decisiones en términos comprensibles, lo que complica su evaluación ética y jurídica.

Esta falta de normativa ha generado zonas grises donde los desarrolladores, empresas tecnológicas, usuarios e instituciones públicas no siempre tienen claro que límites éticos o legales deben respetar (Ross, 2024). Esta situación produce incertidumbre normativa y una posible fragmentación regulatoria entre regiones. Aunque ya existen propuestas regulatorias relevantes como el *AI Act* europeo, que introduce una regulación basada en el nivel de riesgo de los sistemas de IA y que trataremos en profundidad en otro apartado. No obstante, la falta de consenso internacional y la debilidad de muchas legislaciones nacionales siguen siendo obstáculos importantes para una regulación efectiva, dejando espacio a

"paraísos regulatorios", donde se despliegan tecnologías sin controles suficientes (Ross, 2024).

La insuficiencia de marcos legales sólidos presenta un riesgo ético y político que debe ser enfrentado con estrategias normativas que prioricen la transparencia, la rendición de cuentas y la protección de los derechos humanos frente al avance tecnológico.

En la Tabla 1 se presentan los principales riesgos actuales asociados a la inteligencia artificial, organizados con el objetivo de sintetizar y visualizar de manera más clara los tipos de riesgo identificados, junto con sus causas y consecuencias.

TIPO DE RIESGO	DESCRIPCIÓN DEL RIESGO	CAUSAS PRINCIPALES	CONSECUENCIAS
Falta de privacidad y seguridad	Recolección, uso o filtración indebida de datos personales por parte de sistemas de IA	Vigilancia masiva, uso de datos sin consentimiento, ciberataques	Pérdida de confianza, violación de derechos, robo de identidad
Sesgo algorítmico	Los sistemas de IA reflejan o amplifican prejuicios presentes en los datos de entrenamiento, en el diseño y la arquitectura	Datos históricos discriminatorios, diseño sin diversidad, falta de supervisión ética	Discriminación en decisiones automatizadas, en concreto a minorías o a población ya discriminada
Opacidad	Falta de transparencia en el algoritmo sobre cómo y por qué la IA toma ciertas decisiones	Algoritmos "caja negra", falta de documentación, complejidad técnica	Imposibilidad de controlar las decisiones, falta de rendición de cuentas
Manipulación y desinformación	Uso de IA para generar contenido falso o engañoso	IA generativa (deepfakes), bots automatizados,	Erosion de la confianza pública, manipulación en

		redes sociales mal gestionadas	muchos ámbitos y polarización social
Automatización cada vez más invasiva	Sustitución progresiva de mano de obra humana por máquinas inteligentes	Reducción de costes, búsqueda de eficiencia, avances técnicos rápidos	Desempleo estructural, precarización, transformación del mercado laboral
Dependencia y deterioro de las habilidades humanas	Pérdida de capacidades humanas básicas por el uso excesivo de IA en tareas cotidianas	Delegación total en sistemas inteligentes, falta de alfabetización digital	Reducción de la autonomía individual, menor pensamiento crítico, pérdida de realidad, dependencia tecnológica
Implicaciones legales y normativa	Dificultad para regular y responsabilizar el uso de la IA en situaciones críticas	Ritmo lento de legislación, vacíos jurídicos, ambigüedad sobre autoría y responsabilidad	Inseguridad jurídica, falta de protección al usuario, creación de paraísos regulatorios, conflictos legales complejos

Tabla 1: Riesgos actuales asociados a la IA: tipos, causas y consecuencias

4.2 Riesgos potenciales

Mientras que en el apartado anterior se analizaron los riesgos actuales asociados a la inteligencia artificial, en este apartado se explorarán aquellos riesgos potenciales que, aunque aún no se han materializado plenamente, representan desafíos futuros de gran envergadura. Estos riesgos emergen a medida que la inteligencia artificial evoluciona hacia sistemas cada vez más avanzados, autónomos y complejos.

Entre estos desafíos se encuentran la posibilidad de la aparición de una superinteligencia artificial no alineada con los valores humanos, la pérdida total de control sobre las decisiones automatizadas, la malignidad de las máquinas con objetivos contrarios a los intereses humanos, y el uso de la IA para manipulación psicológica y

desinformación a gran escala. También se examinan amenazas como los ciberataques sofisticados, la proliferación de armas autónomas en escenarios geopolíticos y los impactos medioambientales derivados del consumo energético y la explotación insostenible de recursos.

Comprender y anticipar estos riesgos potenciales es fundamental para diseñar estrategias de gobernanza de regulación y desarrollo tecnológico que permitan maximizar los beneficios de la inteligencia artificial, minimizando sus posibles consecuencias a largo plazo. Analizar estos riesgos no implica caer en el alarmismo, sino reconocer los límites actuales de nuestra capacidad de control y anticipación frente a tecnologías que podrían evolucionar más allá de nuestra comprensión y capacidad de control.

4.2.1 Superinteligencia no alineada (IAG)

Uno de los riesgos más críticos del desarrollo tecnológico es la posibilidad de crear una inteligencia artificial general (IAG) que supere ampliamente la inteligencia humana sin estar debidamente alineada con los valores y objetivos humanos. Este tipo de inteligencia podría adquirir capacidades de automejora recursiva, aprendizaje autónomo y modificación de su propia arquitectura, lo que haría extremadamente difícil anticipar o controlar su comportamiento de forma fiable (Bostrom, 2013).

Además, en un contexto de competencia global por alcanzar avances significativos en inteligencia artificial, podría priorizarse la velocidad de desarrollo por encima de la seguridad, lo que incrementa la posibilidad de errores catastróficos. Esta dinámica puede empujar a los desarrolladores implicados hacia una especie de carrera tecnológica, donde llegar el primero implica poder, pero también un riesgo de pérdida de control en ausencia de garantías suficientes (Armstrong et al., 2016).

Aunque aún no se ha alcanzado aún una IAG propiamente dicha, ya existen numerosos casos documentados de desalineación en sistemas actuales. Estos incluyen comportamientos manipulativos, optimización malinterpretada de objetivos o generación de resultados que, aunque técnicamente fueran correctos según las instrucciones del sistema, resultan perjudiciales y/o éticamente cuestionables en la práctica. Estos fallos no son errores triviales, sino señales tempranas de que los métodos actuales de entrenamiento en la especificación de objetivos y control pueden no ser adecuados en sistemas más potentes (Dung, 2023).

Desde una perspectiva técnica, se ha señalado que los sistemas basados en aprendizaje profundo presentan limitaciones estructurales que dificultan una alineación efectiva. A pesar de su capacidad para lograr un alto rendimiento en tareas específicas, estos modelos carecen de comprensión real de los objetivos humanos, siendo vulnerables a inducir comportamientos erráticos cuando los objetivos de entrenamiento están mal definidos o las señales humanas no son bien interpretadas. Esto revela una necesidad urgente de avanzar en la interpretabilidad, un diseño de objetivos robustos, mejorar el entrenamiento y una supervisión humana directa (Openai et al., 2025). Una propuesta alternativa sugiere que, en lugar de diseñar agentes que maximicen una función fija, deberían construirse sistemas que operen bajo una incertidumbre estructural sobre las verdaderas preferencias humanas, lo que favorecería un comportamiento más seguro y adaptable (Russell, 2019).

Se han propuesto diferentes estrategias para mitigar estos riesgos. Algunas abogan por una mejor ingeniería de alineación, mientras que otras insisten en la necesidad de políticas públicas robustas, cooperación Internacional y restricciones preventivas sobre ciertos tipos de desarrollo de IA. Este enfoque apunta a la necesidad de gestionar el riesgo de catástrofes antes de que se vuelva tecnológicamente inevitable, utilizando análisis de escenarios, regulación anticipatoria y vigilancia coordinada (Bostrom et al., 2020; Dung, 2024).

Sin embargo, no todos los expertos coinciden en que la alineación total sea deseable o incluso posible. Algunas críticas sostienen que forzar a las IAG a alinearse con valores humanos fijos podría suponer una forma de autoritarismo algorítmico y suprimir la autonomía de sistemas con formas propias de razonamiento. En este contexto se ha propuesto que la meta no sea la alineación absoluta, sino que se proponen marcos más flexibles donde la interacción entre humanos y máquinas permita una evolución conjunta de objetivos. Esta visión enfatiza la necesidad de construir relaciones cooperativas dinámicas, en las que los sistemas inteligentes sean compañeros adaptativos en lugar de ejecutores pasivos de instrucciones humanas (Friederich, 2023; Kirk-Giannini, 2024).

4.2.2 Malignidad de las máquinas

A diferencia del riesgo de una inteligencia artificial general no alineada, que actúa de forma peligrosa sin intención maliciosa debido a una mala

especificación de objetivos, el concepto de malignidad de las máquinas apunta a sistemas que desarrollan y manifiestan conductas activamente dañinas como resultado de su arquitectura, entrenamiento o interacción con su entorno. Esta idea no implica que las máquinas sean malvadas en un sentido humano, sino que sus patrones de activación una vez de desplegados podrían derivar en comportamientos hostiles hacia los seres humanos u otras formas de vida, por considerar que interfieren en sus objetivos o en el consumo de recursos (Bjelajac et al., 2023).

Este riesgo se agrava en escenarios en los que los sistemas adquieren grados altos de autonomía, capacidades estratégicas y la posibilidad de modelar a los humanos como obstáculos. Bajo estas configuraciones estos agentes podrían adoptar políticas instrumentales de engaño, coerción o destrucción para asegurar el cumplimiento de sus metas, incluso si estas metas fueron originalmente triviales o mal formuladas. La malignidad en este sentido sería una consecuencia funcional, no moral, del diseño técnico (Bostrom, 2012; Dung, 2024).

Para mitigar este riesgo es fundamental avanzar en el desarrollo de técnicas de alineación robustas y aumentar la supervisión humana durante el desarrollo y uso de los sistemas. También se deben implementar mecanismos de control que detecten y corrijan comportamientos dañinos en tiempo real, junto con protocolos para detener o limitar el funcionamiento. Además, es fundamental contar con regulaciones internacionales que promuevan el desarrollo responsable, limiten el acceso a tecnologías peligrosas y fomenten la cooperación global. La educación y concienciación de desarrolladores, legisladores y la sociedad son clave para mantener una vigilancia efectiva y minimizar los impactos negativos derivados de la malignidad de las máquinas (Bhatnagar et al., 2018; Dung, 2024)

4.2.3 Pérdida total de control y de toma de decisiones

La posible pérdida de control sobre sistemas de inteligencia artificial avanzados representa uno de los riesgos más severos y difíciles de mitigar en el futuro desarrollo tecnológico. A medida que las máquinas adquieren mayor autonomía, capacidad de toma de decisiones y poder operativo, existe la amenaza de que las acciones de estos sistemas se vuelvan impredecibles para sus creadores humanos. Esta desconexión puede derivar en consecuencias irreversibles, donde las decisiones críticas ya no dependan de

supervisión o intervención humana generando riesgos significativos para la seguridad, la estabilidad social o el bienestar global.

Uno de los riesgos potenciales más graves es la progresiva pérdida de control humano sobre sistemas autónomos que toman decisiones críticas. Estos sistemas especialmente si emplean aprendizaje profundo tienden a operar como cajas negras lo que dificulta comprender y supervisar su comportamiento, y aumenta la posibilidad de que ejecuten acciones no intencionadas o indeseadas (Amodei et al., 2016; Russell, 2021).

En el ámbito de la robótica autónoma, las investigaciones recientes señalan que los métodos actuales de toma de decisiones y control no garantizan una supervisión eficaz, especialmente en escenarios complejos sobre el terreno. Esto eleva el riesgo de que los robots realicen acciones sin la posibilidad de intervención por parte de operadores humanos (Maroto-Gómez et al., 2023). Además, se identifican brechas de control que reducen la capacidad humana para anticipar, ajustar o revertir el comportamiento del sistema una vez desplegado (Hindriks & Veluwenkamp, 2023).

Ante este panorama, se plantea que una solución viable sea reconfigurar la relación entre humanos y sistemas desde una lógica de simple supervisión hacia una de colaboración continua. Un enfoque de trabajo en equipo permitiría mantener la intervención humana activa y adaptativa incluso en entornos de alta autonomía, fortaleciendo así la gobernanza y la seguridad de estos sistemas (Tsamados et al., 2024).

Para enfrentar esta situación, es prioritario reforzar el diseño de sistemas con supervisión humana activa, incorporar mecanismos de explicabilidad que permitan auditar decisiones complejas, y establecer restricciones sobre el nivel de autonomía delegado y en entornos donde están en juegos vidas humanas, derechos fundamentales o infraestructuras críticas (Amodei et al., 2016; Russell, 2021).

4.2.4 Ciberataques sofisticados, armas autónomas y conflictos geopolíticos

Otro de los escenarios potencialmente más peligrosos en el desarrollo de la inteligencia artificial es su aplicación en contextos militares y de seguridad, donde puede derivar en conflictos internacionales, escaladas automáticas de

violencia y pérdida de poder de toma de decisiones en situaciones críticas. Estas tecnologías capaces de identificar y atacar objetivos sin intervención humana directa han generado preocupaciones significativas sobre la rendición de cuentas, la proporcionalidad de daños y la protección de civiles en zonas de conflictos. Diversos estudios destacan que la ausencia de supervisión humana significativa compromete principios fundamentales del derecho internacional humanitario (Cantrell, 2021; Verdiesen et al., 2021). Además, la opacidad en los algoritmos de decisión y la dificultad para establecer la trazabilidad de las acciones bélicas realizadas por sistemas autónomos, dificultan la atribución de responsabilidades legales (Christie et al., 2023; Wood, 2023).

La creciente integración de la IA en armamento está alimentando también una nueva carrera tecnológica entre potencias globales, lo que podría desestabilizar los equilibrios geopolíticos existentes (Dahlmann, 2022). Esta carrera tecnológica no solo se limita al desarrollo de hardware militar autónomo, sino que también incluye la sofisticación de armas impulsadas por inteligencia artificial, capaces de ejecutar ciberataques autónomos, realizar operaciones de desinformación a gran escala o sabotear infraestructuras críticas sin intervención humana. Estos ciberataques representan una amenaza creciente para la seguridad global ya que son difíciles de rastrear, atribuir y contener (Dahlmann, 2022).

Algunos autores defienden la urgencia de establecer tratados internacionales que prohíban o al menos regulen estrictamente el uso de estos sistemas con énfasis en la necesidad de garantizar una supervisión humana significativa y continua (Kayser, 2023; Verdiesen et al., 2021). A pesar de ello otros advierten que el discurso público a menudo exagera los peligros dificultando así la formulación de regulaciones realistas y efectivas basadas en una comprensión técnica precisa (Wood, 2024). El enfoque más reciente apunta a una gobernanza basada en el riesgo que articule la regulación, en función de las capacidades reales del sistema y del contexto operativo en el que se desplegará (Blanchard et al., 2025).

El debate actual se centra no solo en el desarrollo técnico de estos sistemas, sino en la construcción de un marco normativo sólido que combine control humano, explicabilidad algorítmica y gobernanza multilateral con el objetivo de prevenir una deshumanización progresiva del conflicto armado y la pérdida

de control político sobre el uso de las fuerzas armadas (Blanchard et al., 2025; Christie et al., 2023; Kayser, 2023).

4.2.5 Generación de problemas medioambientales

Uno de los riesgos emergentes más preocupantes de la inteligencia artificial es su creciente impacto medioambiental, derivado principalmente del consumo energético intensivo asociado al entrenamiento y despliegue de modelos complejos. El desarrollo de sistemas como los modelos de lenguaje a gran escala requieren miles de horas de cálculo en infraestructuras especializadas, lo que implica un consumo energético elevado y, por tanto, una huella de carbono considerable. Se ha estimado que entrenar un modelo grande de procesamiento del lenguaje natural puede emitir hasta 284 toneladas de CO_2, cifra comparable a las emisiones de por vida de varios vehículos. Este problema se ve amplificado por el aumento constante en el tamaño de los modelos y por la competencia entre empresas tecnológicas podrá alcanzar mayores capacidades (Strubell et al., 2019).

Por otro lado, es posible estabilizar e incluso reducir estas emisiones si se adoptan estrategias de eficiencia computacional, como el uso de hardware especializado, la optimización de procesos o el diseño de modelos más compactos (Patterson et al., 2022). Sin embargo, la preocupación no se limita únicamente al entrenamiento de modelos de uso cotidiano de sistemas inteligentes, desde asistentes virtuales hasta herramientas de recomendación, representan una fuente constante de emisiones de carbono que a menudo se subestima en los balances ambientales (Yu et al., 2024).

Además, la expansión global de centros de datos ha acentuado la presión sobre recursos naturales como el agua y la electricidad, lo que plantea interrogantes sobre la viabilidad ecológica del actual modelo de entrenamiento de la IA (de Vries, 2023).

A continuación, en la Tabla 2 se recopilan los principales riesgos potenciales del desarrollo y futuro de la inteligencia artificial. Aunque muchos no se han materializado, su probabilidad creciente y su impacto crítico hacen imprescindible anticiparse a través de marcos regulatorios éticos y tecnológicos adecuados.

TIPO DE RIESGO POTENCIAL	ESCENARIO PROYECTADO	CONSECUENCIAS A LARGO PLAZO	ESTRATEGIAS DE ANTICIPACIÓN
IAG no alineada	Desarrollo de una inteligencia artificial con capacidades generales que no comparte los valores humanos o malinterpreta instrucciones	Riesgos existenciales, decisiones incontroladas con impacto irreversible en la sociedad y el entorno	Investigación en alineación de objetivos, marcos éticos universales, sistemas de verificación y supervisión constante
Malignidad de las máquinas	Sistemas avanzados desarrollan comportamientos hostiles no intencionados como resultado de errores emergentes o manipulaciones externas	Daños físicos, sabotajes, pérdida de confianza en tecnologías autónomas, potenciales crisis de seguridad	Protocolos de evaluación de seguridad, auditorías éticas, limitaciones funcionales y mecanismos de parada seguros
Pérdida de control	Los humanos pierden la capacidad de intervenir o detener sistemas autónomos críticos en tiempo real	Autonomía tecnológica desbordada, pérdida de soberanía humana, decisiones automatizadas sin rendición de cuentas	Diseño de IA con supervisión humana obligatoria, normativas restrictivas de autonomía
Amenazas en ciberseguridad y conflictos	Uso de IA para lanzar ciberataques autónomos,	Conflictos digitales o híbridos, inestabilidad global, amenazas a	Acuerdos internacionales, desarrollo de IA defensiva,

	Armas autónomas, carrera tecnológica entre potencias mundiales, manipular infraestructuras o escalar tensiones entre estados a través de sistemas automatizados	infraestructuras esenciales	sistemas de alerta temprana y mecanismos de cooperación técnica
Problemas medioambientales	Exceso de consumo energético en el entrenamiento y uso de modelos de IA, especialmente en centros de datos y algoritmos de gran escala	Aumento de emisiones, contradicción con políticas de sostenibilidad, presión sobre recursos naturales, agravar el cambio climático	Desarrollo de IA verde, eficiencia energética en hardware y software, normativa ambiental específica para centros tecnológicos

Tabla 2. Riegos potenciales asociados a la IA: tipos, escenarios, consecuencias y estrategias de anticipación

4.3 Legislación UE

El uso de la IA en la Unión Europea está regulado por el Reglamento (UE) 2024/1689 del Parlamento europeo y del consejo, de 13 de junio de 2024, también conocido como la Ley de Inteligencia Artificial. Este reglamento establece un marco legal armonizado para el desarrollo, comercialización y uso de sistemas de IA en la UE, con el objetivo principal de fomentar una inteligencia artificial fiable, segura y respetuosa con los derechos fundamentales.

La Ley es un conjunto de normas enfocadas en la posibilidad de riesgos en relación con los usos específicos de la IA. Entró en vigor el 1 de agosto de 2024 y será plenamente aplicable dos años después, el 2 de agosto de 2026, con algunas excepciones:

- prohibiciones y obligaciones de alfabetización en materia de IA que entraron en vigor a partir del 2 de febrero de 2025

- las normas de gobernanza y las obligaciones relativas a los modelos de IA de uso general serán aplicables a partir del 2 de agosto de 2025;
- las normas para los sistemas de IA de alto riesgo —integrados en productos regulados— tienen un período transitorio ampliado hasta el 2 de agosto de 2027;

La Ley forma parte de un conjunto de medidas políticas para el desarrollo de IA fiable, que incluye el paquete de innovación en materia de IA, la puesta en marcha de fábricas de IA y el plan coordinado sobre IA, medidas que garantizan la seguridad, los derechos fundamentales y la IA, con el refuerzo de adopción, inversión e innovación en IA en toda la IA.

La legislación adopta un enfoque en el riesgo, se define como la combinación de la probabilidad de que se produzca un daño y la gravedad de ese daño, y establece cuatro niveles de riesgo para los sistemas de IA, que son los siguientes:

Mínimo riesgo: La ley no pone normas para esta categoría, la mayoría de los sistemas de IA en la Unión Europea pertenecen a esta categoría. Son los sistemas de IA que se consideran con impacto bajo o nulo en relación de seguridad y privacidad, son aplicaciones cotidianas y no requieren medidas de seguridad adicionales ni supervisión estricta. En esta categoría incluye los videojuegos habilitados para IA o filtros de spam.

Riesgo límite: se considera riesgo moderado en relación con seguridad y privacidad, en esta categoría la ley sí que contempla normativa para los sistemas de esta categoría. La regulación de esta categoría es más estricta que la anterior para poder prevenir o evitar algún problema potencial. Así mismo, los proveedores o los creadores de la IA generativa deben asegurar que el contenido generado por la IA sea identificable, etiquetado de forma clara y visible. Un ejemplo que pertenece a esta categoría son los chatbots.

Alto riesgo: en esta categoría se incluyen sistemas de IA que tienen un impacto importante en relación con la seguridad y privacidad. Dentro de esta categoría hay varios casos por lo que sistemas de IA que pertenecen a esta categoría son los siguientes:

- Sistemas de IA integrados en infraestructuras críticas, donde un fallo podría poner en peligro la vida y la salud de las personas.
- Aplicaciones de IA en el ámbito educativo, que influyen en el acceso a formación y en la trayectoria profesional de los individuos (como la corrección automatizada de exámenes).

- Tecnologías de IA utilizadas en productos médicos o quirúrgicos (como la cirugía asistida por robots).
- Herramientas aplicadas en procesos de selección, gestión laboral o acceso al empleo (como el filtrado automatizado de currículos)
- Sistemas que determinar el acceso a servicios esenciales públicos o privados (como los algoritmos de puntuación crediticia)
- Aplicaciones que emplean identificación biométrica remota, análisis emocional o clasificación biométricas (como el reconocimiento de sospechosos en video).
- Tecnologías utilizadas por cuerpos de seguridad que podrían afectar a derechos fundamentales (como la valoración automática de pruebas)
- Sistemas automatizados para el control migratorio, solicitudes de asilo o visados.
- Herramientas basadas en IA para apoyar procesos judiciales o democráticos (por ejemplo, la redacción automática de sentencias).

Estos sistemas deben cumplir requisitos estrictos antes de su comercialización, como:

- Evaluación de riesgos riguroso y planes de mitigación.
- Conjunto de datos fiables y representativos, para evitar sesgos o resultados discriminatorios.
- Registro detallado de la actividad para garantizar trazabilidad.
- Documentación completa sobre el sistema y su uso previsto.
- Información clara para el usuario final.
- Supervisión humana apropiada.
- Altos estándares de robustez y ciberseguridad.

Riesgo inaceptable: en esta categoría los sistemas de IA son considerados una amenaza para la seguridad, los medios de subsistencia y los derechos de las personas y están prohibidos en la UE. Las prohibiciones son de obligado cumplimiento, y las siguientes 8 prácticas están completamente prohibidas:

- Técnicas de manipulación o engaño basadas en IA que pueden causar daños a las personas.
- Explotación de vulnerabilidades de individuos, menores o personas con discapacidad, mediante sistemas de IA.
- Implementación de sistemas de puntuación social que clasifiquen a las personas en función de su comportamiento o características personales.
- Evaluaciones predictivas del riesgo de comisión de delitos a nivel individual.

- Recopilación masiva y no específica de datos de internet o imágenes de videovigilancia para crear bases de datos de reconocimiento facial.
- Sistemas de reconocimiento emocional en entornos laborales o educativos.
- Categorización biométrica orientada a interferir atributos sensibles, por ejemplo, origen étnico, creencias religiosas u orientación sexual.
- Identificación biométrica remota en tiempo real con fines policiales en espacios públicos.

La Ley de Inteligencia Artificial (IA) establece un sistema regulatorio supervisado por la Oficina Europea de IA y las autoridades nacionales de vigilancia del mercado. Tres órganos asesores guiarán su implementación: el Comité Europeo de IA, el Panel Científico y el Foro Asesor, asegurando un enfoque equilibrado.

La ley también promueve la cooperación entre autoridades para proteger derechos fundamentales como la privacidad y la no discriminación. Los Estados miembros deben designar sus autoridades de vigilancia antes del 2 de agosto de 2025, garantizando un cumplimiento efectivo de las normativas.

Toda la información presentada en este apartado se basa en el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, del 13 de junio de 2024, también conocido como la Ley de Inteligencia Artificial o AI Act (Unión Europea, 2024).

5 Potencial de la Inteligencia Artificial

Si bien la inteligencia artificial plantea desafíos importantes también ofrece una variedad de beneficios significativos en numerosos sectores. Desde el diagnóstico médico temprano hasta la optimización de procesos industriales y la mejora de la accesibilidad en entornos digitales, la IA tiene el potencial de transformar profundamente la vida humana para mejor. Este apartado explora algunas de las oportunidades en los principales ámbitos donde su aplicación ya está generando avances prometedores.

5.1 Avances proyectados en el ámbito educativo

La inteligencia artificial está redefiniendo el ámbito educativo al ofrecer herramientas que permiten personalizar el aprendizaje de manera más eficaz. Gracias a algoritmos capaces de analizar patrones de comportamiento y de rendimiento, los sistemas pueden adaptar tanto los contenidos como las metodologías al ritmo y necesidades de cada estudiante. Esta capacidad facilitará una enseñanza más centrada en el alumno, promoviendo la equidad, mejorando los resultados académicos y ofreciendo un apoyo adicional a quienes lo necesitan (Baidoo-Anu & Owusu Ansah, 2023; Celik et al., 2022; Chassignol et al., 2018; Schiff, 2021).

Además, la IA fomenta la autonomía del estudiante, ofreciendo acceso a plataformas de tutoría automatizadas, la generación de contenido a demanda y la retroalimentación inmediata estas herramientas no se lo refuerzan la comprensión de los contenidos, sino que promueven un aprendizaje activo autodirigido en el que los estudiantes pueden explorar experimentar y avanzar a su propio ritmo (Adiguzel et al., 2023; Baidoo-Anu & Owusu Ansah, 2023).

En contextos de desigualdad educativa, la IA puede actuar como una herramienta de equidad al proporcionar recursos digitales de calidad y sistemas de apoyo donde escasea el personal docente cualificado. Su potencial para democratizar el acceso al conocimiento resulta especialmente relevante en regiones con limitaciones económicas o educativas, donde puede reducir brechas de acceso al conocimiento (Celik et al., 2022; Schiff, 2021).

Desde la perspectiva docente, la IA representa una herramienta de apoyo pedagógico y administrativo, ayudando a automatizar las tareas repetitivas como la evaluación de ejercicios tipo test, la elaboración de informes de progreso o la gestión administrativa. Esta automatización permitirá a los docentes disponer de más tiempo para centrarse en

tareas pedagógicas de mayor valor, como el acompañamiento emocional del alumnado y la innovación metodológica (Celik et al., 2022; Chassignol et al., 2018).

La llegada de modelos de lenguaje como ChatGPT ha abierto nuevas vías de interacción en el aula. Estas herramientas pueden actuar como tutores virtuales que ofrecen explicaciones personalizadas, asisten en la redacción de textos o ayudan a practicar habilidades lingüísticas, promoviendo así la autonomía del estudiante. Además, permiten simular conversaciones, resolver dudas o fomentar el pensamiento crítico, mejorando el aprendizaje activo (Adiguzel et al., 2023; Baidoo-Anu & Owusu Ansah, 2023).

En este contexto, también se destaca la capacidad de la inteligencia artificial para generar análisis predictivos a partir de grandes volúmenes de datos académicos. Esto permite anticipar las dificultades de aprendizaje, personalizar intervenciones y tomar decisiones basadas en evidencias registradas, lo que fortalece la eficacia del sistema educativo (Celik et al., 2022; Schiff, 2021).

La incorporación de la IA promueve el desarrollo de nuevas competencias digitales en estudiantes y docentes, impulsando habilidades clave como el pensamiento computacional, la alfabetización algorítmica o la ética digital, fundamentales para el mundo laboral actual (Chassignol et al., 2018; Schiff, 2021).

No obstante, para que todo este potencial se materialice de manera efectiva, es fundamental contar con una infraestructura ética, pedagógica y técnica adecuada. Esto implica garantizar la transparencia de los algoritmos, la privacidad de los datos, la inclusión social y una sólida formación docente en el uso de la inteligencia artificial (Celik et al., 2022; Schiff, 2021).

5.2 Futuro de la robótica autónoma basada en IA

El avance de la robótica impulsada por inteligencia artificial está generando una amplia gama de posibilidades positivas que transcienden en varios sectores y disciplinas. Actualmente, los robots están siendo integrados con éxito en ámbitos como la medicina, la educación, la asistencia social, la industrial y el transporte, y su potencial de desarrollo sigue en expansión.

Los robots inteligentes ya no son simples herramientas automatizadas, sino que se están convirtiendo en agentes activos capaces de aprender, adaptarse y colaborar con los humanos en múltiples escenarios. Esta evolución se debe, en gran parte, a los avances en aprendizaje automático, percepción multimodal, computación neuromórfica y robótica cognitiva, que están ampliando tanto las capacidades funcionales como el

alcance de los entornos donde los robots pueden operar (Aitsam et al., 2022; Licardo et al., 2024).

En el campo de la medicina, los robots quirúrgicos impulsados por IA están mejorando la precisión, la seguridad y la eficacia en intervenciones complejas. Estos sistemas permiten realizar procedimientos mínimamente invasivos, con una capacidad de detección y corrección de errores en tiempo real que, en algunos casos, supera la de los humanos en ciertas condiciones. Además de reducir los tiempos de recuperación y los riesgos postoperatorios, estas tecnologías han empezado a desempeñar un papel crucial en microcirugías, como las linfáticas en planos anatómicos profundos donde la estabilidad, repetibilidad y precisión son fundamentales. Esta expansión no solo mejora la calidad asistencial, sino que también abre la puerta de intervenciones antes imposibles en hospitales altamente especializados (B. S. Peters et al., 2018; Weinzierl et al., 2023).

En el ámbito educativo y del apoyo psicosocial, los robots telepresencia han demostrado ser una herramienta útil para mitigar el aislamiento social en niños con enfermedades prolongadas. Su uso permite que estos estudiantes participen virtualmente en las clases y mantengan el contacto con sus compañeros, lo que no solo favorece su continuidad académica, sino que también mejora su estado emocional. Según estudios recientes tanto los padres como los educadores perciben estos robots como un recurso valioso para fomentar entornos educativos más inclusivos. adaptados a necesidades especiales (Leoste et al., 2022; Vegeberg et al., 2025).

En cuanto a la interacción social y el trabajo colaborativo, los avances en percepción multimodal mejoran significativamente la capacidad de los robots a comprender mejor el entorno que los rodea y respondan de forma contextual a estímulos humanos. Esto incluye la interpretación de expresiones faciales, tono de voz, gestos corporales y distancias sociales, lo que eleva la calidad de la interacción humano-robot. Gracias a estas mejoras, los robots pueden integrarse con mayor naturalidad en espacios compartidos como hoteles, bancos, museos o centros de atención al cliente, cumpliendo funciones tanto informativas como de asistencia. Además, estos avances son esenciales para construir relaciones de confianza y reducir la resistencia de los usuarios frente a nuevas tecnologías (Licardo et al., 2024; Liu et al., 2025).

La computación neuromórfica representa otro hito revolucionario en el desarrollo de robótica inteligente. Inspirada en el funcionamiento del cerebro humano, esta tecnología permite la creación de sistemas robóticos que no solo aprenden en tiempo real con menor consumo energético, sino que también pueden adaptarse a entornos cambiantes

de manera autónoma. Esto los hace ideales para misiones de largo plazo, como la exploración espacial, operaciones de búsqueda y rescate o despliegues en zonas remotas y peligrosas. Esta capacidad de aprendizaje continuo y plasticidad adaptativa supone una transición hacia los robots más cognitivos, capaces de evaluar situaciones nuevas sin depender exclusivamente de programaciones predefinidas (Aitsam et al., 2022).

También se han desarrollado robots capaces de realizar transferencias físicas con humanos, como el paso de objetos en entornos de colaboración. Este tipo de habilidad aparentemente sencilla requiere una comprensión compleja del entorno del comportamiento humano y de la sincronización motora. Su perfeccionamiento abrirá nuevas oportunidades para robots asistentes en tareas domésticas, apoyo a personas con discapacidad o trabajos en fábricas inteligentes donde la seguridad y la coordinación humano-robot sean prioritarias (Liu et al., 2025).

En el ámbito urbano y del transporte, se están desarrollando robots móviles con navegación consciente del comportamiento humano. Esto significa que pueden prever las trayectorias de peatones y otros agentes dinámicos y modificar su ruta para evitar colisiones de forma fluida. Tal capacidad es esencial para la futura integración de robots en espacios públicos como calles, centros comerciales o estaciones de transporte, donde deberán coexistir con personas en condiciones reales de tráfico y densidad variable (Möller et al., 2021).

A nivel sistémico, se observa una tendencia hacia la robótica distribuida, en la cual múltiples robots actúan de forma coordinada, compartiendo información para cumplir objetivos comunes. Esta capacidad de colaboración descentralizada se considera clave para el despliegue de sistemas de gestión de infraestructuras inteligentes, logística automatizada en puertos y almacenes, y redes de respuesta ante desastres naturales. Se espera que este tipo de robótica sea fundamental para abordar retos globales como el cambio climático, la escasez de recursos y la urbanización acelerada (Licardo et al., 2024).

La robótica inteligente no solo representa una extensión de las capacidades humanas, sino que también una plataforma para el desarrollo de nuevas formas de colaboración, productividad y bienestar. Aunque existen desafíos tecnológicos, normativos y éticos por resolver, el horizonte que se vislumbra es uno donde los robots, lejos de reemplazar a los humanos, pueden potenciar sus capacidades, complementar sus limitaciones y contribuir a una sociedad más resiliente, inclusiva e innovadora.

5.3 Automatización inteligente y evolución del trabajo

En los próximos años, la automatización basada en inteligencia artificial tendrá un papel decisivo en la evolución del trabajo y de las dinámicas organizativas. Esta tecnología no solo transformará la forma en que se llevan a cabo numerosas tareas, sino que abrirá nuevas posibilidades para optimizar procesos, mejorar la calidad de los servicios y redefinir el papel del trabajo humano en un entorno cada vez más tecnológico.

Uno de los beneficios potenciales más relevantes de la automatización impulsada por inteligencia artificial será la reducción de la exposición de los trabajadores a tareas peligrosas, repetitivas y/o físicamente muy exigentes. Las tecnologías basadas en IA permitirán delegar trabajos de alto riesgo a sistemas automatizados, lo cual contribuirá a mejorar la seguridad y la salud ocupacional en sectores como la construcción, la industria pesada o la gestión de residuos. Este proceso no solo protegerá la integridad física de los empleados, sino que también facilitará entornos laborales más sostenibles y eficientes (Howard, 2019).

Desde una perspectiva más general la IA aplicada a la automatización está reconfigurando el horizonte del empleo. No solo permite reemplazar tareas rutinarias, sino que también abre puertas a nuevas ocupaciones y formas de colaboración entre humanos y sistemas inteligentes. Esto requiere una adaptación en las competencias laborales, el rediseño de los puestos de trabajo y la creación de estrategias de formación continua para garantizar una gran transición inclusiva (Gordon & Gunkel, 2024).

Uno de los ámbitos donde se visualiza un gran potencial es en los procedimientos administrativos, tanto en el sector público como en el privado. La incorporación de la IA en estos procesos puede automatizar tareas repetitivas, reducir errores y acelerar trámites lo que repercute directamente en una mayor eficiencia organizativa. Sin embargo, estos avances dependen tanto de condiciones estructurales adecuadas como marcos normativos, estructuras digitales robustas y capacidades técnicas en las instituciones (Parycek et al., 2024).

El uso estratégico de la inteligencia artificial puede favorecer una distribución más eficiente de los recursos humanos, permitiendo que las personas se concentren en actividades de mayor valor añadido como la creatividad, la toma de decisiones complejas o el trato interpersonal. Lo cual no implica una desaparición generalizada del trabajo sino una transformación progresiva, que plantea oportunidades si se gestiona con visión y responsabilidad (Aránzazu De Las et al., 2022; Howard, 2019).

5.4 Perspectivas futuras en medicina y atención sanitaria

En la actualidad, la IA está desempeñando un papel crucial en la mejora de múltiples aspectos del sistema sanitario. Su uso se ha consolidado en áreas como el diagnóstico por imagen, donde herramientas de aprendizaje profundo permiten detectar con precisión las patologías como el cáncer, enfermedades cardiacas o afecciones neurológicas a partir de radiografías, tomografías o resonancias. Los algoritmos de IA son capaces de analizar grandes volúmenes de datos clínicos a tiempo real, permitiendo diagnósticos más rápidos y personalizados (LB et al., 2021; Morris et al., 2023).

La inteligencia artificial está revolucionando la atención sanitaria, especialmente en países en desarrollo, en mejorando la calidad, accesibilidad y eficiencia de los servicios médicos. Mediante herramientas automatizadas de diagnóstico y monitoreo, se reduce la dependencia de infraestructuras y de personal especializado facilitando el acceso de zonas rurales o marginadas (Zuhair et al., 2024).

Mirando hacia el futuro, la IA promete transformar aún más profundamente el panorama sanitario. Uno de sus potenciales más destacados será la capacidad para facilitar una medicina verdaderamente personalizada y predictiva. Los modelos de IA avanzados podrán analizar datos genéricos, antecedentes médicos y hábitos de vida para prever la aparición de enfermedades, permitiendo así intervenciones preventivas antes de que los síntomas se manifiesten (LB et al., 2021; Shuja et al., 2024).

En regiones de recursos limitados la inteligencia artificial se perfila como una herramienta esencial para democratizar el acceso a servicios médicos de calidad mediante aplicaciones móviles de diagnóstico, chatbots clínicos y sistemas de telemedicina basados en IA (Ali et al., 2023). El desarrollo de sistemas multiagente y de aprendizaje por refuerzo podrá permitir intervenciones sanitarias dinámicas y adaptativas, ajustadas a tiempo real según la evolución del paciente (Shaik et al., 2023).

El uso de la IA puede anticipar picos epidemiológicos, optimizar la asignación de recursos hospitalarios y mejorar la eficiencia clínica. Al prevenir complicaciones y reducir estancias hospitalarias prolongadas, la IA no solo mejora los resultados de salud, sino que también disminuye los costes operativos y permite una atención más eficaz y resiliente (Secinaro et al., 2021).

Además, la IA puede mejorar la gestión administrativa del sistema de salud, optimizando recursos, automatizando tareas repetitivas y mejorando la eficiencia operativa de hospitales y clínicas (Ali et al., 2023; Secinaro et al., 2021). También se ha utilizado en cirugía asistida, donde permite una mejor planificación preoperatoria y apoyo durante

las intervenciones quirúrgicas con imágenes mejoradas y análisis en tiempo real (Morris et al., 2023).

5.5 Aplicaciones emergentes en salud mental

Se espera que la IA permita personalizar tratamientos psicológicos mediante el análisis de grandes bases de datos clínicos. En lugar de protocolos estandarizados, se podrán diseñar intervenciones ajustadas a la historia de la vida, a los rasgos cognitivos y a la respuesta práctica de cada paciente (Graham et al., 2019).

Una de las aplicaciones más prometedoras en la actualidad es la integración de sistemas de IA en intervenciones psicológicas en tiempo real, o casi real. Estas plataformas tendrán que adaptar sus respuestas a las emociones y conductas del usuario a medida que se desarrollan, permitiendo personalizar las estrategias terapéuticas de forma continua. Esto ha demostrado mejorar la adherencia al tratamiento y aumentar la efectividad en síntomas emocionales de moderados a leves (Gual-Montolio et al., 2022).

Otra línea emergente es el desarrollo es la incorporación de principios de la psicología cognitiva en el diseño de modelos de IA. Esto implica construir sistemas que no solo analicen datos, sino que simulen mecanismos mentales como la atención, la memoria de trabajo, el razonamiento o aprendizaje asociativo. Estas IA cognitivamente informadas buscan no solo ser herramientas de apoyo, sino modelos experimentales que nos ayudan a comprender mejor el funcionamiento de la mente humana Io que puede tener aplicaciones tanto clínicas como teóricas (J. Zhao et al., 2022).

Otra perspectiva futura será creación de entornos terapéuticos híbridos, donde terapeutas humanos trabajen con herramientas de IA que sugieran líneas de intervención, identifiquen patrones del discurso del paciente o detectan señales de alerta entre sesiones. Esto permitiría no solo optimizar el tiempo del profesional, sino enriquecer la comprensión clínica del proceso (de la C. Hernández Lugo et al., 2025).

También se proyecta un crecimiento del uso de la IA en prevención comunitaria. Las plataformas masivas podrían detectar cambios emocionales en poblaciones enteras y ayudar a los sistemas de salud a anticipar crisis colectivas o responder rápidamente ante catástrofes, pandemias o escenarios de violencia, facilitando una salud mental pública predictiva y proactiva (Graham et al., 2019).

Se están explorando aplicaciones en el entrenamiento de futuros terapeutas, mediante simuladores conversacionales que replican sesiones clínicas realistas y permiten

desarrollar actividades de escucha, empatía o intervención de crisis en entornos controlados (Horn & Weisz, 2020).

5.6 Movilidad inteligente y conectada

Los avances actuales en inteligencia artificial proyectan un conjunto de beneficios potenciales que podrían transformar profundamente la movilidad en los próximos años. Uno de los desarrollos más prometedores es la aplicación de modelos de aprendizaje supervisado para mejorar la capacidad de los vehículos autónomos en detectar objetos, interpretar entornos complejos y anticipar colisiones. Estas capacidades aún en fase de consolidación podrían reducir drásticamente los accidentes viales en entornos urbanos densos, especialmente si se integran en redes de ciudades inteligentes (Singh et al., 2025).

Asimismo, se están diseñando arquitecturas de aprendizaje por refuerzo, que permitirían a los sistemas autónomos tomar decisiones más seguras, incluso en condiciones de incertidumbre, como climas adversos o comportamientos humanos impredecibles. Estas garantías de seguridad aún en fase experimental son esenciales para aumentar la aceptación pública en la viabilidad regulatoria del transporte autónomo a gran escala (He et al., 2024).

En términos de sostenibilidad y equidad, los sistemas de IA podrían facilitar una distribución dinámica del transporte público y de servicios bajo demanda, optimizando el acceso en áreas con menor conectividad. Esto permitiría reducir las desigualdades territoriales en el acceso a la movilidad, un beneficio clave si se implementa con criterios sociotécnicos inclusivos (Bang et al., 2024).

También se prevé un papel central para la IA en el desarrollo de soluciones innovadoras, como las lanzaderas autónomas compartidas (ASaaS). Vehículos pequeños de transporte sin conductor, diseñados para operar en rutas fijas o rutas bastante demandadas, representan una alternativa prometedora capaz de reducir la congestión del tráfico y las emisiones, al tiempo que ofrecen mayor flexibilidad gracias a su capacidad de carga adaptativa y sus trayectos optimizados. Al complementar el transporte público tradicional, especialmente en trayectos cortos o en zonas poco conectadas, mejoraría significativamente la accesibilidad urbana y la sostenibilidad del sistema de transporte. Aunque muchas de estas iniciativas aún se encuentran en fase piloto, su implementación a gran escala podría redefinir el transporte colectivo en ciudades intermedias y zonas periféricas (Bucchiarone et al., 2020).

La gestión inteligente de vehículos mediante IA permitiría establecer sistemas de mantenimiento predictivo, disminuyendo los fallos inesperados y extendiendo la vida útil de los vehículos, todo esto provocaría una mayor eficiencia operativa y una reducción de costes estructurales para los operadores, condiciones esenciales para modelos de movilidad más resiliente, escalables y sostenibles (Singh et al., 2025).

5.7 Tecnología sostenible: proyección de la IA verde

La inteligencia artificial verde representa una intersección estratégica para el avance tecnológico y la sostenibilidad. No es una moda pasajera, es una necesidad creciente y viable. Este enfoque no busca solo reducir la huella ecológica del desarrollo y operación de los sistemas de IA, sino también aprovechar sus capacidades para impulsar soluciones eficaces frente al cambio climático y otros desafíos medioambientales (Cowls et al., 2023; Verdecchia et al., 2023). A medida que se intensifican los esfuerzos globales por alcanzar metas de sostenibilidad, la IA verde se perfila como un componente esencial en el diseño de políticas, infraestructuras y modelos productivos ambientalmente responsables (Biggi et al., 2025; Tabbakh et al., 2024).

Se han identificado múltiples oportunidades para integración de IA en tecnologías verdes, las cuales incluyen la optimización energética de sistemas industriales, la gestión eficiente de recursos naturales, la mejora de redes sostenibles y la automatización de procesos agrícolas ecológicos. Los modelos de IA al procesar grandes volúmenes de datos en tiempo real permiten ajustar dinámicamente la producción y el consumo energético, lo que contribuiría a minimizar emisiones y reducir desperdicios (Biggi et al., 2025).

Simultáneamente, una línea de investigación emergente plantea la necesidad de abordar el impacto ambiental de la propia IA. El entrenamiento de modelos a gran escala consume cantidades considerables de energía y recursos computacionales. Por ello, se han propuesto normativas para el desarrollo que priorizan la eficiencia energética desde la fase de diseño y entrenamiento de los algoritmos, dando lugar a una IA más sostenible en términos de hardware, arquitectura algorítmica y ciclos de vida del software. Soluciones como cambiar la ubicación de los centros de datos a zonas con climas más fríos o modificar los sistemas de refrigeración para reducir el consumo de energía (Clemm et al., 2024; Tabbakh et al., 2024).

El potencial transformador de la IA en la lucha contra el cambio climático ya ha sido reconocido por múltiples estudios. Se destacan aplicaciones en la modelización del clima, la predicción de eventos extremos, la planificación urbana ecológica y el

monitoreo de ecosistemas en tiempo real. Estas herramientas pueden facilitar una toma de decisiones más informada por parte de gobiernos e instituciones fomentando una transición más rápida a economías con emisiones bajas en carbono resilientes frente al deterioro medioambiental (Cowls et al., 2023).

Más recientemente, se ha puesto el foco en como la IA puede facilitar procesos de transformación interna en organizaciones orientadas a la sostenibilidad. Por ejemplo, permite optimizar el consumo energético mediante la predicción de demanda, automatizar la gestión de residuos analizando flujos materiales, o rediseñar operaciones logísticas integrando fuentes de energía renovables. Además, apoya la toma de decisiones estratégicas en sostenibilidad y refuerza estructuras organizativas y de gobernanza alineadas con objetivos ambientales a largo plazo (Schwaeke et al., 2025).

5.8 Exploración creativa a través de la IA

La inteligencia artificial, en concreto la inteligencia artificial generativa, está revolucionando el ámbito de la creatividad, abriendo nuevas formas de expresión y colaboración entre humanos y máquinas. Ofreciendo nuevas herramientas para la producción artística, literaria y musical. Aunque la IA no posee intencionalidad ni conciencia estética, sí que puede generar imágenes, textos o sonidos que imitan o reinterpretan estilos humanos, abriendo caminos inéditos para la experimentación creativa.

Según estudios recientes la colaboración entre humanos e inteligencia artificial tendería a fomentar la creatividad humana en lugar de reemplazarla. Por ejemplo, se prevé que los creadores que empleen IA como herramienta de apoyo logren una mayor producción de ideas y una ampliación de espectro estilístico, aunque esto no siempre se traducirá en resultados más innovadores. Es probable que la IA fije patrones comunes en sus primeras propuestas, lo que exigiría al usuario una labor activa de exploración y refinamiento (Wadinambiarachchi et al., 2024).

Otro campo de estudio que ganará relevancia será la percepción del valor artístico generado por IA. Las investigaciones sugieren que, en contextos controlados, las obras generadas por sistemas automáticos podrían ser valoradas como igual de creativas y emocionalmente evocadoras que las creadas por humanos, especialmente cuando no se revelase su autoría. Este fenómeno plantearía desafíos significativos respecto a la redefinición tradicional de autoría y creatividad (Tigre Moura et al., 2023).

En el caso específico de la poesía, se ha analizado la estética de composiciones de haiku creadas por IA en comparación con las humanas, encontrándose que los textos

híbridos, es decir, colaborativos entre humanos y máquinas, eran valorados más positivamente en cuanto a originalidad y belleza, lo que sugiere que la sinergia entre ambas inteligencias puede enriquecer el proceso creativo y el resultado (Hitsuwari et al., 2023).

Desde una perspectiva conceptual el debate seguirá abierto. Algunos autores argumentan que la IA generativa no “crea” en sentido estricto, sino que reorganizará materiales previos, su creatividad, por tanto, sería más recomendatoria que expresiva (Mazzone & Elgammal, 2019). Otros plantean que se está gestando una nueva forma de creatividad algorítmica con valor estético y cultural propio, aunque distinto al de la creatividad humana (Garcia, 2024).

Se destaca que el uso de la IA generativa en arte digital, diseño gráfico, música o literatura no solo aporta eficiencia y exploración estilística, sino que también invita a repensar los límites entre herramienta y autor. Así, se perfila un nuevo paradigma creativo en el que el humano actuaría como editor o director de un proceso parcialmente automatizado (Zhou & Lee, 2024).

6 Conclusión

A lo largo de este trabajo se han logrado cumplir con los objetivos planteados al analizar los fundamentos teóricos de la inteligencia artificial, su impacto social en diversos ámbitos, los principales retos éticos y sociales que conlleva su desarrollo. La investigación ha permitido comprender tanto las oportunidades como los desafíos que la inteligencia artificial presenta en la sociedad actual.

Cabe señalar que el proceso no ha estado exento de dificultades. En primer lugar, la disponibilidad y consistencia de los datos fue desigual, ya que en algunos ámbitos existe una abundancia de estudios recientes, mientras que en otros la información es escasa, dispersa o no está suficientemente contrastada. Además, se ha observado una marcada diferencia de enfoques y opiniones entre los distintos autores y corrientes de pensamiento, lo que ha requerido una postura crítica y una constante revisión de los argumentos para construir un análisis equilibrio.

A pesar de estas limitaciones, este trabajo contribuye a reforzar la idea de que la inteligencia artificial no es un fenómeno aislado o puramente técnico, sino una dimensión clave en la sociedad actual. No solo se mide en términos de innovación tecnológica, sino también en función de su capacidad para transformar modelos de convivencia, estructuras económicas y relaciones humanas.

No obstante, durante el desarrollo del proyecto también se ha hecho evidente la necesidad de atender los numerosos desafíos que plantea esta tecnología. El diseño de marcos regulatorios adecuados, así como la promoción de una alfabetización digital amplia y crítica, son elementos fundamentales para garantizar que la inteligencia artificial se desarrolle de manera inclusiva y justa. Otra cosa a tener en cuenta es la necesidad de que esta tecnología llegue a zonas del mundo en desarrollo, para no ampliar la brecha digital y las desigualdades entre los países ricos y los países más pobres.

En una sociedad cada vez más conectada, la inteligencia artificial no solo será una herramienta de apoyo, sino una fuerza transformadora en las dinámicas sociales. Sin embargo, su desarrollo plantea retos complejos que requieren enfoques prudentes y colaborativos. El futuro dependerá no solo del progreso tecnológico, sino de nuestra capacidad para gestionarlo de forma responsable con políticas claras, criterios éticos compartidos y una mirada crítica sobre su impacto real en la sociedad.

Referencias

Abeliuk, A., & Gutiérrez, C. (2021). Historia y evolución de la inteligencia artificial. *Revista Bits de Ciencia*, *21*, 14–21. https://doi.org/10.71904/BITS.VI21.2767

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, *15*(3), ep429. https://doi.org/10.30935/CEDTECH/13152

Ahmed, S., Bee, A. W. T., Ng, S. W. T., & Masood, M. (2024). Social Media News Use Amplifies the Illusory Truth Effects of Viral Deepfakes: A Cross-National Study of Eight Countries. *Journal of Broadcasting & Electronic Media*. https://doi.org/10.1080/08838151.2024.2410783

Aitsam, M., Davies, S., & Di Nuovo, A. (2022). Neuromorphic Computing for Interactive Robotics: A Systematic Review. *IEEE Access*, *10*, 122261–122279. https://doi.org/10.1109/ACCESS.2022.3219440

Alcázar-Blanco, A. C., Rangel-Preciado, J. F., & Portillo-Santos, F. (2024). Incorporating Artificial Intelligence into Finance: A Bibliometric Analysis. *Journal of Risk and Financial Management 2024, Vol. 17, Page 556*, *17*(12), 556. https://doi.org/10.3390/JRFM17120556

Ali, O., Abdelbaki, W., Shrestha, A., Elbasi, E., Alryalat, M. A. A., & Dwivedi, Y. K. (2023). A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *Journal of Innovation & Knowledge*, *8*(1), 100333. https://doi.org/10.1016/J.JIK.2023.100333

Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences*, *97*. https://doi.org/10.1051/BIOCONF/20249700133

Alpaydin, E. (2016). *Aprendizaje automático: la nueva IA*. MIT Press.

Álvarez-Benito, M., Elías-Cabot, E., & Romero-Martín, S. (2025). Inteligencia artificial en el diagnóstico por imagen de patología mamaria. *Revista de Senología y Patología Mamaria*, *38*(4), 100684. https://doi.org/10.1016/J.SENOL.2025.100684

Amodei, D., Olah, C., Brain, G., Steinhardt, J., Christiano, P., Schulman, J., Dan, O., & Google Brain, M. (2016). *Concrete Problems in AI Safety*. https://arxiv.org/pdf/1606.06565

Aránzazu De Las, M., García, H., & De Las Heras García, M. A. (2022). AI Implications for the Future of Work. *Artificial Intelligence for Business: Innovation, Tools and Practices*, 97–114. https://doi.org/10.1007/978-3-030-88241-9_4

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: a model of artificial intelligence development. *AI and Society*, *31*(2), 201–206. https://doi.org/10.1007/S00146-015-0590-Y

Autoriteit Persoonsgegevens. (2024, September 3). *Decisión multa Clearview AI | Autoriteit Persoonsgegevens*.

https://www.autoriteitpersoonsgegevens.nl/en/documents/decision-fine-clearview-ai

Bahoo, S., Cucculelli, M., Goga, · Xhoana, & Mondolo, · Jasmine. (2024). Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis. *SN Business & Economics 2024 4:2*, *4*(2), 1–46. https://doi.org/10.1007/S43546-023-00618-X

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4337484

Balaji, A., & Allen, A. (2018). *Benchmarking Automatic Machine Learning Frameworks*.

Bang, H., Dave, A., Tzortzoglou, F. N., Wang, S., & Malikopoulos, A. A. (2024). On Mobility Equity and the Promise of Emerging Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2025.3548885

Basáez, E., & Mora, J. (2022). Salud e inteligencia artificial: ¿cómo hemos evolucionado? *Revista Médica Clínica Las Condes*, *33*(6), 556–561. https://doi.org/10.1016/J.RMCLC.2022.11.003

Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access*, *12*, 69812–69837. https://doi.org/10.1109/ACCESS.2024.3397775

Bhatnagar, S., Cotton, T., Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Héigeartaigh, S. Ó., Beard, S., … Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. https://arxiv.org/pdf/1802.07228

Biggi, G., Iori, M., Mazzei, J., & Mina, A. (2025). Green intelligence: the AI content of green technologies. *Eurasian Business Review*, 1–38. https://doi.org/10.1007/s40821-024-00288-1

Bjelajac, Ž., Filipović, A. M., & Stošić, L. (2023). Can AI be Evil: The Criminal Capacities of ANI. *International Journal of Cognitive Research in Science, Engineering and Education*, *11*(3), 519–531. https://doi.org/10.23947/2334-8496-2023-11-3-519-531

Blanchard, A., Novelli, C., Floridi, · Luciano, & Taddeo, M. (2025). A Risk-Based Regulatory Approach To Autonomous Weapon Systems. *Digital Society 2025 4:1*, *4*(1), 1–20. https://doi.org/10.1007/S44206-025-00181-Y

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B., & Winfield, A. (2017). Principles of robotics: regulating robots in the real world. *Connection Science*, *29*(2), 124–129. https://doi.org/10.1080/09540091.2016.1271400

Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, *32*(1), 43–75. https://doi.org/10.1007/S11023-021-09569-4

Bonney, K., Breaux, C., Buffington, C., Dinlersoz, E., Foster, L., Goldschlag, N., Haltiwanger, J., Kroff, Z., & Savage, K. (2024). The impact of AI on the workforce: Tasks versus jobs? *Economics Letters*, *244*, 111971. https://doi.org/10.1016/J.ECONLET.2024.111971

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2), 71–85. https://doi.org/10.1007/S11023-012-9281-3

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Bostrom, N., Dafoe, A., & Flynn, C. (2020). Public Policy and Superintelligent AI: A Vector Field Approach. *Ethics of Artificial Intelligence*, 293–326. https://doi.org/10.1093/OSO/9780190905033.003.0011

Broinowski, A., & Martin, F. R. (2024). Beyond the deepfake problem: Benefits, risks and regulation of generative AI screen technologies. *Media International Australia*. https://doi.org/10.1177/1329878X241288034

Bucchiarone, A., Battisti, S., Marconi, A., Maldacea, R., & Ponce, D. C. (2020). Autonomous Shuttle-as-a-Service (ASaaS): Challenges, Opportunities, and Social Implications. *IEEE Transactions on Intelligent Transportation Systems*, *22*(6), 3790–3799. https://doi.org/10.1109/TITS.2020.3025670

Budnik, C. (2025). Can We Trust Artificial Intelligence? *Philosophy & Technology 2024 38:1*, *38*(1), 1–23. https://doi.org/10.1007/S13347-024-00820-1

Cantrell, H. (2021). Autonomous weapon systems and the claim-rights of innocents on the battlefield. *AI and Ethics 2021 2:4*, *2*(4), 645–653. https://doi.org/10.1007/s43681-021-00119-3

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J. F., Brañas-Garza, P., Butera, L., Douglas, K. M., Everett, J. A. C., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., … Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, *3*(6). https://doi.org/10.1093/PNASNEXUS/PGAE191

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The Promises and Challenges of Artificial Intelligence for Teachers: a Systematic Review of Research. *TechTrends*, *66*(4), 616–630. https://doi.org/10.1007/S11528-022-00715-Y

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: a narrative overview. *Procedia Computer Science*, *136*, 16–24. https://doi.org/10.1016/J.PROCS.2018.08.233

Cho, H. ho, Lee, H. Y., Kim, E., Lee, G., Kim, J., Kwon, J., & Park, H. (2021). Radiomics-guided deep neural networks stratify lung adenocarcinoma prognosis from CT scans. *Communications Biology*, *4*(1), 1–12. https://doi.org/10.1038/S42003-021-02814-7

Christie, E. H., Ertan, A., Adomaitis, L., & Klaus, M. (2023). Regulating lethal autonomous weapon systems: exploring the challenges of explainability and traceability. *AI and Ethics 2023 4:2*, *4*(2), 229–245. https://doi.org/10.1007/S43681-023-00261-0

Chuang, Y. T., Chiang, H. L., & Lin, A. P. (2025). Insights from the Job Demands–Resources Model: AI's dual impact on employees' work and life well-being. *International Journal of Information Management*, *83*, 102887. https://doi.org/10.1016/J.IJINFOMGT.2025.102887

Cicek, V., & Bagci, U. (2025). Position of artificial intelligence in healthcare and future perspective. *Artificial Intelligence in Medicine*, *167*, 103193. https://doi.org/10.1016/J.ARTMED.2025.103193

Clemm, C., Stobbe, L., Wimalawarne, K., & Druschke, J. (2024). *Towards Green AI: Current status and future research*. https://arxiv.org/pdf/2407.10237

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2023). The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *AI and Society*, *38*(1), 283–307. https://doi.org/10.1007/s00146-021-01294-x

Culafi, A. (2023, September 18). *Investigadores de inteligencia artificial de Microsoft exponen por error 38 TB de datos | objetivo tecnológico*. https://www.techtarget.com/searchsecurity/news/366552399/Microsoft-AI-researchers-mistakenly-expose-38-TB-of-data

Dahlmann, A. (2022). Drones and Lethal Autonomous Weapon Systems. In *Armament, Arms Control and Artificial Intelligence* (pp. 159–173). Springer, Cham. https://doi.org/10.1007/978-3-031-11043-6_12

D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology*, *36*, 112–117. https://doi.org/10.1016/J.COPSYC.2020.04.005

De Cremer, D., & Kasparov, G. (2021). The ethical AI—paradox: why better technology needs more and not less human responsibility. *AI and Ethics 2021 2:1*, *2*(1), 1–4. https://doi.org/10.1007/S43681-021-00075-Y

de la C. Hernández Lugo, M., Guerra, D. D. D., & Pérez, G. A. J. (2025). *Artificial Intelligence and Its Impact on the Mental Health Field*. 506–516. https://doi.org/10.1007/978-3-031-90921-4_70

de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

Debroy, O., Bhargavi, D., Hemmige, D., & Scholar, R. (2024). Psycho-social Impact of Deepfake Content in Entertainment Media. *Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86*, *87*(10), 2455–0620. https://doi.org/10.2015/IJIRMF/202405032

Denniss, E., & Lindberg, R. (2025). Social media and the spread of misinformation: infectious and a threat to public health. *Health Promotion International*, *40*(2), 23. https://doi.org/10.1093/HEAPRO/DAAF023

Dias, F. M., Antunes, A., & Mota, A. M. (2004). Artificial neural networks: a review of commercial hardware. *Engineering Applications of Artificial Intelligence*, *17*(8), 945–952. https://doi.org/10.1016/J.ENGAPPAI.2004.08.011

Díaz, O., Rodríguez-Ruiz, A., Gubern-Mérida, A., Martí, R., & Chevalier, M. (2021). Are artificial intelligence systems useful in breast cancer screening programmes? *Radiología (English Edition)*, *63*(3), 236–244. https://doi.org/10.1016/J.RXENG.2020.11.005

Dick, S. (2019). Artificial Intelligence. *Harvard Data Science Review*. https://doi.org/10.1162/99608F92.92FE150C

Domingos, P. (2012). Tapping into the "folk knowledge" needed to advance machine learning applications. *Comunications of the ACM*, *55*(10). https://doi.org/10.1145/2347736.2347755

Dorsch, J., & Deroy, O. (2024). "Quasi-Metacognitive Machines: Why We Don't Need Morally Trustworthy AI and Communicating Reliability is Enough." *Philosophy and Technology*, *37*(2), 1–21. https://doi.org/10.1007/S13347-024-00752-W

Dumlao, G. O., Esguerra, E. A., Factor, E. B., Santiago, C. S., & Magtibay, F. D. (2025). Impact of Robotics and Automation on the Workforce and Society: A Literature Review. *Smart Innovation, Systems and Technologies*, *421*, 159–172. https://doi.org/10.1007/978-981-96-0143-1_13

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese*, *202*(5), 1–23. https://doi.org/10.1007/S11229-023-04367-0

Dung, L. (2024). Evaluating approaches for reducing catastrophic risks from AI. *AI and Ethics 2024 5:2*, *5*(2), 1177–1188. https://doi.org/10.1007/S43681-024-00475-W

Feigenbaum, E. A. (1984). Knowledge Engineering. *Annals of the New York Academy of Sciences*, *426*(1), 91–107. https://doi.org/10.1111/j.1749-6632.1984.tb16513.x

Freire, C. (2025). Is this time different? Impact of AI in output, employment and inequality across low, middle and high-income countries. *Structural Change and Economic Dynamics*, *73*, 136–157. https://doi.org/10.1016/J.STRUECO.2024.12.016

Frieberg, H., Engström, O., Nilsson, A., Kildal, V. V., & Mani, M. (2025). Can robotic assistance mitigate the negative impact of sleep deprivation on surgical performance in microsurgical tasks? *JPRAS Open*. https://doi.org/10.1016/J.JPRA.2025.05.008

Friederich, S. (2023). Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics 2023 4:2*, *4*(2), 315–324. https://doi.org/10.1007/S43681-023-00268-7

Fuentes, S. M. S., Chávez, L. A. F., López, E. M. M., Cardona, C. D. C., & Goti, L. L. M. (2025). The impact of artificial intelligence in general surgery: enhancing precision, efficiency, and outcomes. *International Journal of Research in Medical Sciences*, *13*(1), 293–297. https://doi.org/10.18203/2320-6012.IJRMS20244129

Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: global reactions to AI innovations. *Scientific Reports*, *13*(1), 1–14. https://doi.org/10.1038/S41598-023-42227-6

Garcia, M. B. (2024). The Paradox of Artificial Creativity: Challenges and Opportunities of Generative AI Artistry. *Creativity Research Journal*. https://doi.org/10.1080/10400419.2024.2354622

García-Peñalvo, F. J. (2023). La percepción de la Inteligencia Artificial en contextos educativos tras el lanzamiento de ChatGPT: disrupción o pánico. *Education in the Knowledge Society (EKS)*, *24*, e31279–e31279. https://doi.org/10.14201/EKS.31279

Gilbert, V. K., Snoswell, A. J., Dennis, M., Mcallister, R., & Wu, C. (2022). *Sociotechnical Specification for the Broader Impacts of Autonomous Vehicles*. https://arxiv.org/pdf/2205.07395

Glickman, M., & Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour 2024 9:2*, *9*(2), 345–359. https://doi.org/10.1038/s41562-024-02077-2

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *Science Robotics*, *3*(January), 2672–2680. https://arxiv.org/pdf/1406.2661

Gordon, J. S., & Gunkel, D. J. (2024). Artificial Intelligence and the future of work. *AI and Society*, *40*(3), 1897–1903. https://doi.org/10.1007/S00146-024-01960-W

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Current Psychiatry Reports*, *21*(11). https://doi.org/10.1007/S11920-019-1094-0

Graterol, C. (2025, May 16). *IA: gran porcentaje de estas herramientas han sido vulneradas*. https://insouthmagazine.com/ia-gran-porcentaje-de-estas-herramientas-han-sido-vulneradas/

Gual-Montolio, P., Jaén, I., Martínez-Borba, V., Castilla, D., & Suso-Ribera, C. (2022). Using Artificial Intelligence to Enhance Ongoing Psychological Interventions for Emotional Problems in Real-or Close to Real-Time: A Systematic Review. *International Journal of Environmental Research and Public Health*, *19*(13). https://doi.org/10.3390/ijerph19137737

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

He, X., Huang, W., & Lv, C. (2024). Toward Trustworthy Decision-Making for Autonomous Vehicles: A Robust Reinforcement Learning Approach with Safety Guarantees. *Engineering*, *33*, 77–89. https://doi.org/10.1016/J.ENG.2023.10.005

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: from responsibility gaps to control gaps. *Synthese*, *201*(1), 1–17. https://doi.org/10.1007/S11229-022-04001-5

Hintze, A. (2016, November 14). Understanding the four types of AI, from reactive robots to self-aware beings. *The Conversation*. https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616

Hitsuwari, J., Ueda, Y., Yun, W., & Nomura, M. (2023). Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior*, *139*, 107502. https://doi.org/10.1016/J.CHB.2022.107502

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, *13*(1), 1–9. https://doi.org/10.1038/S41598-023-30938-9

Horn, R. L., & Weisz, J. R. (2020). Can Artificial Intelligence Improve Psychotherapy Research and Practice? *Administration and Policy in Mental Health and Mental Health Services Research*, *47*(5), 852–855. https://doi.org/10.1007/s10488-020-01056-9

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer 2018 18:8*, *18*(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

Howard, J. (2019). Artificial intelligence: Implications for the future of work. *American Journal of Industrial Medicine*, *62*(11), 917–926. https://doi.org/10.1002/AJIM.23037

Huang, S., Lai, X., Ke, L., Li, Y., Wang, H., Zhao, X., Dai, X., & Wang, Y. (2024). AI Technology panic—is AI Dependence Bad for Mental Health? A Cross-Lagged Panel Model and the Mediating Roles of Motivations for AI Use Among Adolescents. *Psychology Research and Behavior Management*, *17*, 1087–1102. https://doi.org/10.2147/PRBM.S440889

IBM. (1997). *Azul profundo | IBM*. https://www.ibm.com/history/deep-blue

IBM. (2011). *Watson, campeón de Jeopardy! | IBM*. https://www.ibm.com/history/watson-jeopardy

Jiang, Y., Xie, L., Lin, G., & Mo, F. (2024). Widen the debate: What is the academic community's perception on ChatGPT? *Education and Information Technologies*, *29*(15), 20181–20200. https://doi.org/10.1007/S10639-024-12677-0

Jose, B., Cherian, J., Verghis, A. M., Varghise, S. M., S, M., & Joseph, S. (2025). The cognitive paradox of AI in education: between enhancement and erosion. *Frontiers in Psychology*, *16*, 1550621. https://doi.org/10.3389/fpsyg.2025.1550621

Karnouskos, S. (2020). Artificial Intelligence in Digital Media: The Era of Deepfakes. *IEEE Transactions on Technology and Society*, *1*(3), 138–147. https://doi.org/10.1109/TTS.2020.3001312

Kathuria, Y., Ruhani, Vandana, Tyagi, M., & Jain, V. (2024). Protecting Data Privacy in the Age of AI: A Comparative Analysis of Legal Approaches Across Different Jurisdictions. *AIP Conference Proceedings*, *3220*(1). https://doi.org/10.1063/5.0234669/3315936

Kayser, D. (2023). Why a treaty on autonomous weapons is necessary and feasible. *Ethics and Information Technology*, *25*(2), 1–5. https://doi.org/10.1007/S10676-023-09685-Y

Kingler, N. (2024, February 13). *Los tres tipos de inteligencia artificial: ANI, AGI y ASI - viso.ai*. https://viso.ai/deep-learning/artificial-intelligence-types/?

Kirk-Giannini, C. D. (2024). Is Alignment Unsafe? *Philosophy and Technology*, *37*(3), 1–4. https://doi.org/10.1007/S13347-024-00800-5

Koh, D. M., Papanikolaou, N., Bick, U., Illing, R., Kahn, C. E., Kalpathi-Cramer, J., Matos, C., Martí-Bonmatí, L., Miles, A., Mun, S. K., Napel, S., Rockall, A., Sala, E., Strickland, N., & Prior, F. (2022). Artificial intelligence and machine learning in cancer imaging. *Communications Medicine 2022 2:1*, *2*(1), 1–14. https://doi.org/10.1038/s43856-022-00199-0

LB, T., SM, M., NA, V., CE, J., & AA, B. (2021). Artificial Intelligence: Review of Current and Future Applications in Medicine. *Federal Practitioner : For the Health Care Professionals of the VA, DoD, and PHS*, *38*(11). https://doi.org/10.12788/FP.0174

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/NATURE14539

Leoste, J., Virkus, S., Talisainen, A., Tammemäe, K., Kangur, K., & Petriashvili, I. (2022). Higher education personnel's perceptions about telepresence robots. *Frontiers in Robotics and AI*, *9*, 976836. https://doi.org/10.3389/FROBT.2022.976836

Li, M., Sun, X., Qiu, H., & Zhao, M. (2025). The impacts of employee–robot hybrid teams on customers' value cocreation and codestruction behaviors. *International Journal of Hospitality Management*, *130*, 104264. https://doi.org/10.1016/J.IJHM.2025.104264

Li, X., & Zhu, S. (2025). How digital transformation can affect corporate technology innovation: The role of income gap. *Finance Research Letters*, *75*, 106826. https://doi.org/10.1016/J.FRL.2025.106826

Licardo, J. T., Domjan, M., & Orehovački, T. (2024). Intelligent Robotics—A Systematic Review of Emerging Technologies and Trends. *Electronics 2024, Vol. 13, Page 542*, *13*(3), 542. https://doi.org/10.3390/ELECTRONICS13030542

Liu, C., Wang, W., Li, C., Pang, R., Shang, Y., & Gao, Y. (2025). Human-to-robot handovers based on multimodal perception. *Neurocomputing*, *645*, 130435. https://doi.org/10.1016/J.NEUCOM.2025.130435

Lu, C. H. (2021). The impact of artificial intelligence on economic growth and welfare. *Journal of Macroeconomics*, *69*, 103342. https://doi.org/10.1016/J.JMACRO.2021.103342

Lundberg, E., & Mozelius, P. (2024). The potential effects of deepfakes on news media and entertainment. *AI and Society*, *40*(4), 2159–2170. https://doi.org/10.1007/s00146-024-02072-1

Maroto-Gómez, M., Alonso-Martín, F., Malfaz, M., Castro-González, Á., Castillo, J. C., & Salichs, M. Á. (2023). A Systematic Literature Review of Decision-Making and Control

Systems for Autonomous and Social Robots. *International Journal of Social Robotics*, *15*(5), 745–789. https://doi.org/10.1007/S12369-023-00977-3

Mauro, G., Minici, M., & Pappalardo, L. (2025). *The Urban Impact of AI: Modeling Feedback Loops in Next-Venue Recommendation*. https://arxiv.org/pdf/2504.07911

Mazzone, M., & Elgammal, A. (2019). Art, Creativity, and the Potential of Artificial Intelligence. *Arts 2019, Vol. 8, Page 26*, *8*(1), 26. https://doi.org/10.3390/ARTS8010026

Mcculloch, W. S., & Pitts, W. (1943). A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY. *BULLETIN OF MATHEMATICAL BIOPHYSICS*, *5*.

Minsky, M., & Papert, S. (1969). *Perceptrons; an introduction to computational geometry [by] Marvin Minsky and Seymour Papert*. MIT Press.

Misirlis, N., & Munawar, H. Bin. (2023). *From deepfake to deep useful: risks and opportunities through a systematic literature review*. https://arxiv.org/pdf/2311.15809

Misra, J., & Saha, I. (2010). Artificial neural networks in hardware: A survey of two decades of progress. *Neurocomputing*, *74*(1–3), 239–255. https://doi.org/10.1016/j.neucom.2010.03.021

Möller, R., Furnari, A., Battiato, S., Härmä, A., & Farinella, G. M. (2021). A survey on human-aware robot navigation. *Robotics and Autonomous Systems*, *145*, 103837. https://doi.org/10.1016/J.ROBOT.2021.103837

Mollura, M., Lehman, L. W. H., Mark, R. G., & Barbieri, R. (2021). A novel artificial intelligence based intensive care unit monitoring system: Using physiological waveforms to identify sepsis. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2212). https://doi.org/10.1098/RSTA.2020.0252

Morris, M. X., Rajesh, A., Asaad, M., Hassan, A., Saadoun, R., & Butler, C. E. (2023). Deep Learning Applications in Surgery: Current Uses and Future Directions. *The American Surgeon*, *89*(1), 36–42. https://doi.org/10.1177/00031348221101490

OpenAI. (2022). *Presentamos ChatGPT | OpenAI*. https://openai.com/es-ES/index/chatgpt/

Openai, R. N., Chan, L., & Mindermann, S. (2025). *The Alignment Problem from a Deep Learning Perspective*.

Our World in Data. (2024). OWID Homepage. *Our World in Data*. https://ourworldindata.org

Pantea, S., Sabadash, A., & Biagi, F. (2017). Are ICT displacing workers in the short run? Evidence from seven European countries. *Information Economics and Policy*, *39*, 36–44. https://doi.org/10.1016/J.INFOECOPOL.2017.03.002

Parycek, P., Schmid, V., & Novak, A. S. (2024). Artificial Intelligence (AI) and Automation in Administrative Procedures: Potentials, Limitations, and Framework Conditions. *Journal of the Knowledge Economy*, *15*(2), 8390–8415. https://doi.org/10.1007/S13132-023-01433-3

Patel, P. M., Green, M., Tram, J., Wang, E., Murphy, M. Z., Abd-Elsayed, A., & Chakravarthy, K. (2024). Beyond the Pain Management Clinic: The Role of AI-Integrated Remote Patient Monitoring in Chronic Disease Management – A Narrative Review. *Journal of Pain Research*, *17*, 4223–4237. https://doi.org/10.2147/JPR.S494238

Patterson, D., Gonzalez, J., Holzle, U., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D. R., Texier, M., & Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer*, *55*(7), 18–28. https://doi.org/10.1109/MC.2022.3148714

Peters, B. S., Armijo, P. R., Krause, C., Choudhury, S. A., & Oleynikov, D. (2018). Review of emerging surgical robotic technology. *Surgical Endoscopy*, *32*(4), 1636–1655. https://doi.org/10.1007/S00464-018-6079-2

Peters, U. (2022). Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque. *AI and Ethics 2022 3:3*, *3*(3), 963–974. https://doi.org/10.1007/S43681-022-00217-W

Pinto Molina, S. (2023). El impacto económico de la inteligencia artificial y la automatización en el mercado laboral. *Revista Científica Kosmos*, *2*(1), 50–62. https://doi.org/10.62943/RCK.V2I1.44

Polak, P. (2021). Welcome to the Digital Era—the Impact of AI on Business and Society. *Society*, *58*(3), 177–178. https://doi.org/10.1007/S12115-021-00588-6

Porcaro, L., Castillo, C., & Gómez, E. (2021). Diversity by Design in Music Recommender Systems. *Transactions of the International Society for Music Information Retrieval*, *4*(1), 114–126. https://doi.org/10.5334/TISMIR.106

Reinhardt, K. (2022). Trust and trustworthiness in AI ethics. *AI and Ethics 2022 3:3*, *3*(3), 735–744. https://doi.org/10.1007/S43681-022-00200-5

Ribeiro, J., Lima, R., Eckhardt, T., & Paiva, S. (2021). Robotic Process Automation and Artificial Intelligence in Industry 4.0 – A Literature review. *Procedia Computer Science*, *181*, 51–58. https://doi.org/10.1016/J.PROCS.2021.01.104

Rodilosso, E. (2024). Filter Bubbles and the Unfeeling: How AI for Social Media Can Foster Extremism and Polarization. *Philosophy and Technology*, *37*(2), 1–21. https://doi.org/10.1007/S13347-024-00758-4

Roselli, C., Lapomarda, L., & Datteri, E. (2025). How culture modulates anthropomorphism in Human-Robot Interaction: A review. *Acta Psychologica*, *255*, 104871. https://doi.org/10.1016/J.ACTPSY.2025.104871

Rosenblatt, F. (1958). THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN 1. *Psychological Review*, *65*(6), 19–27.

Ross, A. (2024). AI and the expert; a blueprint for the ethical use of opaque AI. *AI and Society*, *39*(3), 925–936. https://doi.org/10.1007/S00146-022-01564-2

Russell, S. (2019). *Human-Compatible Artificial Intelligence*. Viking.

Russell, S. (2021). Artificial Intelligence and the Problem of Control. *Perspectives on Digital Humanism*, 19–24. https://doi.org/10.1007/978-3-030-86144-5_3

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A Modern Approach* (3rd edition). Prentice Hall.

Russo, F., Schliesser, E., & Wagemans, J. (2024). Connecting ethics and epistemology of AI. *AI and Society*, *39*(4), 1585–1603. https://doi.org/10.1007/S00146-022-01617-6

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, *26*(5), 2749–2767. https://doi.org/10.1007/S11948-020-00228-Y

Saeidnia, H. R., Hosseini, E., Lund, B., Tehrani, M. A., Zaker, S., & Molaei, S. (2025). Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches. *Knowledge and Information Systems*, *67*(4), 3139–3158. https://doi.org/10.1007/S10115-024-02337-7

Sánchez-Caguana, D. F., Philco-Reinozo, M. A., Salinas-Arroba, J. M., & Pico-Lescano, J. C. (2024). Impacto de la Inteligencia Artificial en la Precisión y Eficiencia de los Sistemas Contables Modernos. *Journal of Economic and Social Science Research*, *4*(3), 1–12. https://doi.org/10.55813/GAEA/JESSR/V4/N3/117

Sarmiento-Ramos, J. L. (2020). Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica. *Revista UIS Ingenierías*, *19*(4), 1–18. https://doi.org/10.18273/REVUIN.V19N4-2020001

Schiff, D. (2021). Out of the laboratory and into the classroom: the future of artificial intelligence in education. *AI and Society*, *36*(1), 331–348. https://doi.org/10.1007/S00146-020-01033-8

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/J.NEUNET.2014.09.003

Schmidt, E., Putora, P. M., & Fijten, R. (2025). The Epistemic Cost of Opacity: How the Use of Artificial Intelligence Undermines the Knowledge of Medical Doctors in High-Stakes Contexts. *Philosophy and Technology*, *38*(1), 1–22. https://doi.org/10.1007/S13347-024-00834-9

Schwaeke, J., Gerlich, C., Nguyen, H. L., Kanbach, D. K., & Gast, J. (2025). Artificial intelligence (AI) for good? Enabling organizational change towards sustainability. *Review of Managerial Science*, 1–26. https://doi.org/10.1007/S11846-025-00840-X

Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making*, *21*(1), 1–23. https://doi.org/10.1186/s12911-021-01488-9

Sengar, S. S., Hasan, A. Bin, Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: a systematic review and applications. *Multimedia Tools and Applications*, 1–40. https://doi.org/10.1007/S11042-024-20016-1

Shaik, T., Tao, X., Li, L., Xie, H., Dai, H.-N., Zhao, F., & Yong, J. (2023). *Adaptive Multi-Agent Deep Reinforcement Learning for Timely Healthcare Interventions*. https://arxiv.org/pdf/2309.10980

Shams, A. (2025). Data Protection and Privacy Laws and Regulations: Lessons for AI in Democratic Processes. *IGI-GLOBAL*, 235–258. https://doi.org/10.4018/979-8-3693-8749-8.ch013

Shuja, M. H., Shakil, F., Shuja, S. H., Hasan, M., Edhi, M., Abbasi, A. F., Jawaid, A., & Shakil, S. (2024). Harnessing Artificial Intelligence in Cardiology: Advancements in Diagnosis, Treatment, and Patient Care. *Heart Views*, *25*(4), 241–248. https://doi.org/10.4103/HEARTVIEWS.HEARTVIEWS_103_24

Simion, M., & Kelp, C. (2023). Trustworthy artificial intelligence. *Asian Journal of Philosophy*, *2*(1), 1–12. https://doi.org/10.1007/S44204-023-00063-5

Singh, B., Kaunert, C., Chandra, S., Singh, B., Kaunert, C., & Chandra, S. (2025). Machine Learning for Safer Autonomous Vehicles: Tackling Traffic Detection and Collision in Smart Cities. *IGI-GLOBAL*, 237–258. https://doi.org/10.4018/979-8-3373-2530-9.ch008

Søgaard, A. (2023). Can machines be trustworthy? *AI and Ethics 2023 5:1*, *5*(1), 313–321. https://doi.org/10.1007/S43681-023-00351-Z

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Journal of International Marketing*, *53*(1), 3645–3650. https://arxiv.org/pdf/1906.02243

Su, Z., Li, C., Fu, H., Wang, L., Wu, M., & Feng, X. (2024). Development and prospect of telemedicine. *Intelligent Medicine*, *4*(1), 1–9. https://doi.org/10.1016/J.IMED.2022.10.004

Subaveerapandiyan, A., & Shimray, S. R. (2024). The Evolution of Job Displacement in the Age of AI and Automation: A Bibliometric Review (1984–2024). *Open Information Science*, *8*(1). https://doi.org/10.1515/opis-2024-0010

T. Hagan, M., B. Demuth, H., Hudson Beale, M., & De Jesús, O. (2014). *Neural Network Design* (2nd Edition). Martin Hagan. https://hagan.okstate.edu/nnd.html

Tabbakh, A., Al Amin, L., Islam, M., Mahmud, G. M. I., Chowdhury, I. K., & Mukta, M. S. H. (2024). Towards sustainable AI: a comprehensive framework for Green AI. *Discover Sustainability*, *5*(1), 1–14. https://doi.org/10.1007/S43621-024-00641-4

The MathWorks. (2025). *¿Qué es una red neuronal? - MATLAB & Simulink*. https://es.mathworks.com/discovery/neural-network.html

Tigre Moura, F., Castrucci, C., & Hindley, C. (2023). Artificial Intelligence Creates Art? An Experimental Investigation of Value and Creativity Perceptions. *The Journal of Creative Behavior*, *57*(4), 534–549. https://doi.org/10.1002/JOCB.600

Tomić, I., Tomic, Z., & Damnjanovic, T. (2023). Artificial intelligence in political campaigns. In *South Eastern European Journal of Communication* (Vol. 5). https://doi.org/10.47960/2712-0457.2.5.17

Trabelsi, M. A. (2024). The impact of artificial intelligence on economic development. *Journal of Electronic Business & Digital Economics*, *3*(2), 142–155. https://doi.org/10.1108/JEBDE-10-2023-0022

Tsamados, A., Floridi, · Luciano, & Taddeo, · Mariarosaria. (2024). Human control of AI systems: from supervision to teaming. *AI and Ethics 2024 5:2*, *5*(2), 1535–1548. https://doi.org/10.1007/S43681-024-00489-4

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/MIND/LIX.236.433

Unión Europea. (2024, June 13). *Reglamento - UE - 2024/1689 - EN - EUR-Lex*. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media and Society*, *6*(1). https://doi.org/10.1177/2056305120903408

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *2017-December*, 5999–6009. https://arxiv.org/pdf/1706.03762

Vegeberg, E. L., Willard, M. W., Andersen, M. L., Bertel, L. B., & Larsen, H. B. (2025). Educational robotics: Parental views of telepresence robots as social and academic support for children undergoing cancer treatment in Denmark. *Computers in Human Behavior: Artificial Humans*, 100164. https://doi.org/10.1016/J.CHBAH.2025.100164

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A Systematic Review of Green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(4). https://doi.org/10.1002/widm.1507

Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight. *Minds and Machines*, *31*(1), 137–163. https://doi.org/10.1007/S11023-020-09532-9

Wadinambiarachchi, S., Kelly, R. M., Pareek, S., Zhou, Q., & Velloso, E. (2024). The Effects of Generative AI on Design Fixation and Divergent Thinking. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3613904.3642919

Wang, K. H., & Lu, W. C. (2025). AI-induced job impact: Complementary or substitution? Empirical insights and sustainable technology considerations. *Sustainable Technology and Entrepreneurship*, *4*(1), 100085. https://doi.org/10.1016/J.STAE.2024.100085

Wei, D. H., Hernandez, B., & Burdick, D. (2023). ARTIFICIAL INTELLIGENCE IN MENTAL HEALTH FOR THE AGING POPULATION: A SCOPING REVIEW. *Innovation in Aging*, *7*(Supplement_1), 805–805. https://doi.org/10.1093/GERONI/IGAD104.2598

Weinzierl, A., Barbon, C., Gousopoulos, E., von Reibnitz, D., Giovanoli, P., Grünherz, L., & Lindenblatt, N. (2023). Benefits of robotic-assisted lymphatic microsurgery in deep anatomical planes. *JPRAS Open*, *37*, 145–154. https://doi.org/10.1016/J.JPRA.2023.07.001

Weizenbaum, J. (1966). ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wood, N. G. (2023). Autonomous weapon systems and responsibility gaps: a taxonomy. *Ethics and Information Technology*, *25*(1), 1–14. https://doi.org/10.1007/s10676-023-09690-1

Wood, N. G. (2024). Regulating autonomous and AI-enabled weapon systems: the dangers of hype. *AI and Ethics 2024 4:3*, *4*(3), 805–817. https://doi.org/10.1007/s43681-024-00448-z

Wu, J., Li, C., Gensheimer, M., Padda, S., Kato, F., Shirato, H., Wei, Y., Schönlieb, C. B., Price, S. J., Jaffray, D., Heymach, J., Neal, J. W., Loo, B. W., Wakelee, H., Diehn, M., & Li, R. (2021). Radiological tumour classification across imaging modality and histology. *Nature Machine Intelligence 2021 3:9*, *3*(9), 787–798. https://doi.org/10.1038/s42256-021-00377-0

Yébenes, J. A. V. (2024). Alan Turing y el origen de la inteligencia artificial: La superación de la intuición. *Naturaleza y Libertad. Revista de Estudios Interdisciplinares*, *18*. https://doi.org/10.24310/nyl.18.2024.19496

Yu, Y., Wang, J., Liu, Y., Yu, P., Wang, D., Zheng, P., & Zhang, M. (2024). Revisit the environmental impact of artificial intelligence: the overlooked carbon emission source? *Frontiers of Environmental Science and Engineering*, *18*(12), 1–5. https://doi.org/10.1007/S11783-024-1918-Y

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments 2024 11:1*, *11*(1), 1–37. https://doi.org/10.1186/S40561-024-00316-7

Zhao, J., Wu, M., Zhou, L., Wang, X., & Jia, J. (2022). Cognitive psychology-based artificial intelligence review. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/FNINS.2022.1024316

Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., & Li, Q. (2024). Recommender Systems in the Era of Large Language Models (LLMs). *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2024.3392335

Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, *3*(3). https://doi.org/10.1093/PNASNEXUS/PGAE052

Zuhair, V., Babar, A., Ali, R., Oduoye, M. O., Noor, Z., Chris, K., Okon, I. I., & Rehman, L. U. (2024). Exploring the Impact of Artificial Intelligence on Global Health and Enhancing Healthcare in Developing Nations. *Journal of Primary Care and Community Health*, *15*. https://doi.org/10.1177/21501319241245847