Por Roberto Segundo Tejada Rodriguez, Santiago Gonzales Mesia, José Luis Castro Ullilen, Juan José Palomino Ochoa, Rosario Leonor Palomino Ochoa, María Micaela Castillo De Lima, Claudia Patricia Yon Delgado

Inteligencia artificial: un enfoque moderno y aprendizaje profundo

ISBN: 978-9915-698-71-7

9 789915 698717

www.editorialmarcaribe.es

EST. 2021 EMC
EDITORIAL MAR CARIBE

Inteligencia artificial: un enfoque moderno y aprendizaje profundo

Tejada Rodriguez, Roberto Segundo; Gonzales Mesia, Santiago; Castro Ullilen, José Luis; Palomino Ochoa, Juan José; Palomino Ochoa, Rosario Leonor; Castillo De Lima, María Micaela; Yon Delgado, Claudia Patricia

Primera edición (1.ª ed.): febrero, 2026

Editado por:

Editorial Mar Caribe ®

www.editorialmarcaribe.es

Av. Gral. Flores 547, 70000 Col. del Sacramento, Departamento de Colonia, Uruguay.

Diseño de carátula e ilustraciones: *Luisa Fernanda Lugo Rojas*

Libro electrónico disponible en:

https://editorialmarcaribe.es/ark:/10951/isbn.9789915698717

Formato: Electrónico

ISBN: 978-9915-698-71-7

ARK: ark:/10951/isbn.9789915698717

Editorial Mar Caribe (OASPA): Como miembro de la Open Access Scholarly Publishing Association, apoyamos el acceso abierto de acuerdo con el código de conducta, la transparencia y las mejores prácticas de OASPA para la publicación de libros académicos y de investigación. Estamos comprometidos con los más altos estándares editoriales en ética y deontología, bajo la premisa de «Ciencia Abierta en América Latina y el Caribe»

Editorial Mar Caribe, firmante Nº 795 de 12.08.2024 de la Declaración de Berlín

"... Nos sentimos obligados a abordar los retos de Internet como medio funcional emergente para la distribución del conocimiento. Obviamente, estos avances pueden modificar significativamente la naturaleza de la publicación científica, así como el actual sistema de garantía de calidad...." (Max Planck Society, ed. 2003, pp. 152-153).

Editorial Mar Caribe se adhiere a la *"Recomendación relativa a la preservación del patrimonio documental, comprendido el patrimonio digital, y el acceso al mismo"* de la UNESCO y a la Norma Internacional de referencia para un sistema abierto de información archivística (OAIS-ISO 14721). Este libro está preservado digitalmente por ARAMEO.NET

Editorial Mar Caribe

Inteligencia artificial: un enfoque moderno y aprendizaje profundo

Colonia, Uruguay

2026

Inteligencia artificial: un enfoque moderno y aprendizaje profundo

Índice

Introducción

Al considerar el estado de la disciplina en el primer trimestre de 2026, resulta evidente que la inteligencia artificial (IA) ha trascendido su identidad histórica como mera herramienta de automatización para consolidarse como un socio cognitivo fundamental en la infraestructura de la civilización contemporánea. El paradigma ha virado drásticamente desde la interrogante clásica planteada a mediados del siglo XX sobre si las máquinas podrían algún día pensar, hacia una indagación mucho más profunda y centrada en la condición humana: de qué manera el pensamiento, la creatividad y la toma de decisiones éticas de las personas evolucionan al operar de forma integrada con sistemas inteligentes.

Esta transformación no es solo técnica, sino ontológica, alterando la percepción de la capacidad humana en campos tan diversos como la medicina de precisión, la ciencia de materiales y la gobernanza global. La arquitectura de este libro surge de la necesidad de sintetizar dos vertientes que, hasta hace poco, se trataban como dominios paralelos: el enfoque moderno de agentes racionales y la profundidad técnica de las redes neuronales contemporáneas.

Ahora bien, el enfoque moderno ya no puede definirse únicamente por la maximización de funciones de rendimiento estáticas en entornos controlados; hoy, la modernidad en IA implica la creación de sistemas que poseen una comprensión estructurada y auditable del mundo, capaces de razonar sobre la incertidumbre y de justificar sus decisiones ante supervisores humanos y marcos regulatorios estrictos. El aprendizaje profundo, por su

parte, ha evolucionado desde el procesamiento masivo de datos hacia una fase de eficiencia extrema y especialización, lo que permite que modelos con un menor número de parámetros superen en tareas específicas a los gigantes monolíticos de años anteriores.

Esta investigación se sitúa en un momento en el que la IA ha dejado de ser un instrumento externo para convertirse en un compañero que eleva el potencial humano en el trabajo y en la resolución de problemas. En el ámbito de la investigación científica, por ejemplo, la IA ya no se limita a resumir la literatura académica, sino que participa activamente en el proceso de descubrimiento en física, química y biología, generando hipótesis y controlando experimentos de forma autónoma, pero bajo la gobernanza humana. Esta capacidad de agencia es el hilo conductor que recorre los capítulos de esta obra, con el objetivo de analizar cómo el aprendizaje profundo proporciona la percepción necesaria para una dimensión sociotécnica que trasciende la definición clásica de agentes racionales que operan en entornos estáticos.

La modernidad hoy se define por la integración de la IA en procesos estratégicos donde la explicación, la defensa y el refinamiento de las decisiones son tan importantes como el resultado mismo. Para comprender la sofisticación de los sistemas actuales, es imperativo analizar la trayectoria evolutiva que comenzó con las visiones proféticas de los pioneros de la computación. Alan Turing, en su trabajo fundamental de 1950, no solo propuso el famoso test de comportamiento inteligente, sino que también estableció la base teórica de la máquina universal, un dispositivo capaz de ejecutar cualquier cómputo descriptible mediante un conjunto de instrucciones. No

obstante, las raíces de la inteligencia artificial simbólica se extienden aún más atrás, hasta Ada Lovelace, quien vislumbró que la Máquina Analítica podría procesar algo más que números, manipulando símbolos que representaran música o texto, lo que sentó las bases del procesamiento simbólico moderno.

Capítulo I

Inteligencia artificial: un enfoque moderno y aprendizaje profundo

El panorama contemporáneo de la inteligencia artificial representa la culminación de siete décadas de investigación y de milenios de pensamiento filosófico sobre la naturaleza del razonamiento y la acción. En la transición hacia la tercera década del siglo XXI, el campo ha experimentado un giro paradigmático desde la ingeniería del conocimiento manual hacia un modelo centrado en los datos, en el que la arquitectura de los agentes inteligentes se define por su capacidad para extraer jerarquías conceptuales a partir de la experiencia sensorial directa. Este informe examina la síntesis entre el enfoque de agentes racionales, popularizado por Stuart Russell y Peter Norvig, y la revolución del aprendizaje profundo liderada por Ian Goodfellow, Yoshua Bengio y Aaron Courville, y analiza cómo dicha convergencia está dando lugar a sistemas neurosimbólicos capaces de emular tanto la intuición rápida como el razonamiento lógico deliberado.

1.1 El marco del agente inteligente y la evolución de AIMA

La definición de inteligencia artificial ha convergido en torno al estudio de los agentes que reciben percepciones del entorno y ejecutan acciones para maximizar una medida de desempeño. Este enfoque, detallado en la cuarta edición de *Artificial Intelligence: A Modern Approach* (AIMA), unifica subcampos previamente fragmentados, como la lógica, la probabilidad y el cálculo

continuo, bajo un marco común centrado en la toma de decisiones óptima (Russel & Norvig, 2004).

A diferencia de las ediciones anteriores, la versión más reciente de AIMA refleja la madurez del aprendizaje automático y su impacto en la robótica, la visión por computadora y el procesamiento del lenguaje natural. Un cambio filosófico crítico es el abandono de la suposición de que el objetivo del agente es fijo y conocido por el sistema; ahora se asume que el agente puede tener incertidumbre sobre los verdaderos objetivos humanos, lo que requiere un aprendizaje continuo de las preferencias del usuario para operar de manera segura y ética.

La estructura del texto se ha reorganizado para reflejar la prevalencia de los datos masivos y de los recursos de computación modernos. Aproximadamente el 25% del material es completamente nuevo y el 75% ha sido reescrito para integrar conceptos como la programación probabilística, los sistemas multiagente y, de manera prominente, el aprendizaje profundo como componente central del diseño de agentes.

1.2 Estructura y taxonomía del entorno de la tarea (PEAS)

Para diseñar un agente, primero se debe especificar el entorno de la tarea utilizando el marco PEAS, que comprende la Medida de Desempeño (Performance), el Entorno (Environment), los Actuadores (Actuators) y los Sensores (Sensors). Este análisis permite determinar qué tipo de arquitectura es necesaria (Russel & Norvig, 2004).

La racionalidad del agente se evalúa en función de su capacidad para

seleccionar acciones que maximicen el valor esperado de la medida de desempeño, dada la secuencia de percepciones recibidas hasta el momento y el conocimiento previo incorporado. Es fundamental distinguir entre racionalidad y omnisciencia: un agente racional toma la mejor decisión posible con la información disponible, aunque el resultado final no sea óptimo debido a factores imprevisibles (Frankish & Ramsey, 2014).

Tipos de arquitecturas de agentes

La complejidad del agente está intrínsecamente ligada a la naturaleza de su entorno. Los agentes se clasifican en varias categorías según su estructura interna y su capacidad de procesamiento:

1. Agentes de Reflejo Simple: Actúan basándose únicamente en la percepción actual, mediante reglas de condición-acción. Son limitados en entornos no totalmente observables.
2. Agentes de Reflejo Basados en Modelos: Mantienen un estado interno que representa las partes no visibles del mundo y se actualizan según el historial de percepciones.
3. Agentes Basados en Metas: Utilizan información sobre objetivos deseables para planificar secuencias de acciones. Introducen la necesidad de la búsqueda y de la planificación.
4. Agentes Basados en Utilidad: Emplean una función de utilidad para cuantificar el grado de felicidad o de éxito de un estado, lo que permite tomar decisiones cuando hay objetivos en conflicto o incertidumbre.
5. Agentes de Aprendizaje: Incorporan un componente que permite mejorar el desempeño a través de la experiencia, ajustando los elementos de

decisión en función de la retroalimentación del entorno.

Algoritmos clásicos de búsqueda y planificación

A pesar del auge del aprendizaje profundo, los algoritmos clásicos de búsqueda siguen siendo la base de la planificación y el razonamiento lógico en sistemas de IA. Estos se dividen principalmente en búsquedas no informadas e informadas.

Búsqueda no informada y heurística

Los métodos de búsqueda no informada, como la búsqueda en anchura (BFS) y la búsqueda en profundidad (DFS), exploran el espacio de estados sin conocer la proximidad de la meta. El algoritmo de Dijkstra, o búsqueda de costo uniforme, expande el nodo con el menor costo acumulado, garantizando la optimalidad si los costos son positivos.

La búsqueda informada utiliza funciones heurísticas para guiar el proceso. El algoritmo A* es el estándar en este ámbito, combinando el costo acumulado $g(n)$ con una estimación del costo restante $h(n)$ para encontrar el camino óptimo de manera eficiente. En entornos adversarios, como los juegos de suma cero, se emplean el algoritmo Minimax y la poda Alfa-Beta para reducir drásticamente el espacio de búsqueda al descartar ramas que no afectan la decisión final.

Representación del conocimiento y planificación

La IA moderna distingue entre tres tipos de representación de estado: atómica (el estado es una caja negra), factorizada (el estado se describe por

atributos) y estructurada (el estado incluye objetos y sus relaciones). La planificación automatizada, detallada en los capítulos de AIMA, utiliza estas representaciones para realizar búsquedas en el espacio de estados, ya sea hacia adelante (desde el estado inicial hasta la meta) o hacia atrás (desde la meta hasta el estado inicial) (Tejada et al., 2017).

1.3 El paradigma del aprendizaje profundo (Deep Learning)

El aprendizaje profundo ha transformado la capacidad de las máquinas para procesar datos no estructurados como imágenes, audio y texto. Se define como una forma de aprendizaje automático que permite a las computadoras aprender de la experiencia a través de una jerarquía de conceptos. Al construir conceptos complejos a partir de otros más simples, los sistemas de aprendizaje profundo eliminan la necesidad de que los humanos definan manualmente todas las características relevantes de los datos.

El éxito de las redes neuronales profundas se apoya en tres áreas matemáticas críticas: el álgebra lineal para la representación de tensores, la teoría de la probabilidad para modelar la incertidumbre y el cálculo para la optimización mediante el descenso de gradiente estocástico. El libro fundamental de Goodfellow et al. divide el campo en tres partes: conceptos matemáticos básicos, redes neuronales prácticas modernas y perspectivas de investigación avanzada (ver Tabla 1).

Tabla 1: Función en el aprendizaje profundo de las matemáticas

Área Matemática	**Función en el Aprendizaje Profundo**
Álgebra Lineal	Operaciones con matrices y vectores para transformar datos.
Probabilidad e Información	Cuantificación de la incertidumbre y medición de la sorpresa en los datos.
Computación Numérica	Optimización iterativa de pesos para minimizar la función de pérdida.
Cálculo Multivariable	Cálculo de los gradientes para la retropropagación del error.

Las redes neuronales convolucionales (CNN) son la arquitectura de referencia para el procesamiento de imágenes y videos. Utilizan capas de convolución que aplican filtros para detectar características espaciales de forma jerárquica: desde bordes y texturas en las capas iniciales hasta objetos complejos en las capas profundas.

Por otro lado, las redes neuronales recurrentes (RNN) y sus variantes avanzadas, como LSTM (Long Short-Term Memory) y GRU (Gated Recurrent Unit), están diseñadas para procesar datos secuenciales. Estas redes mantienen un estado oculto que actúa como memoria, lo que permite que la información de pasos temporales previos influya en las predicciones actuales. Sin embargo, las RNN enfrentan desafíos con las dependencias a largo plazo debido al problema del desvanecimiento del gradiente.

El auge de los Transformers y la atención

La arquitectura Transformer, introducida en 2017, ha reemplazado en gran medida a las RNN en tareas de procesamiento de lenguaje natural (NLP). A diferencia de las RNN, los Transformers procesan secuencias completas en paralelo, utilizando un mecanismo de atención para ponderar la relevancia de cada parte de la entrada respecto de las demás. Este paralelismo no solo reduce el tiempo de entrenamiento, sino que también permite capturar dependencias de largo alcance de manera mucho más efectiva.

Los Transformers utilizan la codificación posicional para mantener el orden de las palabras, ya que el mecanismo de atención es inherentemente agnóstico de la posición. Esta arquitectura es la base de modelos de lenguaje a gran escala (LLM) como GPT-4 y Gemini, que han demostrado capacidades emergentes en razonamiento y generación de contenido.

Integración neuro-simbólica: El futuro de la IA

A pesar de los avances en el aprendizaje profundo, los modelos neuronales puros a menudo carecen de interpretabilidad y sufren de opacidad, operando como cajas negras. La inteligencia artificial neuro-simbólica surge como un enfoque híbrido que combina la robustez perceptiva del aprendizaje profundo con la capacidad de razonamiento lógico de la IA simbólica clásica (Liang et al., 2025).

Esta síntesis se alinea con la teoría del pensamiento rápido y lento de Daniel Kahneman. El Sistema 1 (representado por las redes neuronales) es rápido, intuitivo y experto en el reconocimiento de patrones, pero propenso a

errores lógicos. El Sistema 2 (representado por la IA simbólica) es lento, deliberativo y lógico, capaz de seguir reglas explícitas y realizar deducciones rigurosas.

La IA neuro-simbólica integra ambos sistemas para crear modelos que no solo reconozcan una imagen de un accidente, por ejemplo, sino que también comprendan las leyes de tránsito y puedan explicar por qué se produjo el evento.

Arquitecturas y Logic Tensor Networks (LTN)

Existen diversas formas de integrar componentes neuronales y simbólicos, desde envoltorios simbólicos que filtran las salidas de una red hasta sistemas en los que la lógica está profundamente integrada en la estructura de la red. Las Logic Tensor Networks (LTN) representan un avance significativo al permitir que las fórmulas de lógica de primer orden se traduzcan en términos diferenciables en una red neuronal.

En una LTN, las reglas de conocimiento experto se incorporan directamente en la función de pérdida. Esto permite que el sistema aprenda patrones en los datos mientras respeta restricciones lógicas absolutas. Por ejemplo, en el diagnóstico médico, una red neuronal puede detectar patrones en una radiografía, mientras que un componente simbólico garantiza que el diagnóstico sugerido no contradiga principios biológicos fundamentales ni pautas clínicas establecidas.

AlphaGeometry: Un hito de razonamiento híbrido

AlphaGeometry, un sistema desarrollado por Google DeepMind,

ejemplifica el poder de la IA neurosimbólica al resolver problemas de geometría de alto nivel. Su arquitectura consta de dos partes principales (Wang et al., 2025):

- Un modelo de lenguaje neuronal que predice construcciones auxiliares (nuevos puntos o líneas) que podrían resultar útiles para la prueba.
- Un motor de deducción simbólica que aplica reglas lógicas de geometría clásica para avanzar en la demostración.

Comparativa: IA basada en reglas vs. Aprendizaje profundo

La elección entre un sistema basado en reglas y uno basado en aprendizaje profundo depende de la complejidad de la tarea y de la necesidad de transparencia y control (ver Tabla 2).

Tabla 2: Comparativa: IA basada en reglas vs. Aprendizaje profundo

Criterio	**Sistemas basados en Reglas**	**Aprendizaje Profundo**
Explicabilidad	Alta: Cada decisión se rastrea hasta una regla.	Baja: funciona como una caja negra.
Requerimientos de Datos	Bajos o nulos: depende de expertos humanos.	Muy altos: requieren millones de puntos de datos.
Manejo de Incertidumbre	Bajo: Las reglas suelen ser deterministas.	Alto: Excelente para manejar el ruido y los matices.
Escalabilidad	Difícil: El número de reglas crece exponencialmente.	Moderada: escala en potencia de cómputo y de datos.

Adaptabilidad	Requiere intervención manual constante.	Alta: Aprende y se ajusta con nuevos datos.

La tendencia actual no es elegir uno frente a otro, sino fusionarlos en sistemas que mantengan la precisión y la adaptabilidad del aprendizaje profundo, mientras conservan la auditabilidad y el cumplimiento de las normas de los sistemas expertos.

Ecosistema de la IA en Lima: Recursos y formación

La adopción de la inteligencia artificial en el Perú, y específicamente en Lima, ha impulsado la creación de una infraestructura educativa y de investigación significativa. El distrito de Jesús María y sus alrededores se han convertido en un eje para el acceso a estos recursos técnicos.

Instituciones académicas y programas especializados

Varias universidades en Lima han lanzado carreras y laboratorios enfocados exclusivamente en IA para preparar a la próxima generación de profesionales.

1. Universidad María Auxiliadora (UMA): Ofrece la carrera de Ingeniería en Inteligencia Artificial, con cursos modulares que incluyen Deep Learning, Reinforcement Learning e IA para Ciberseguridad. Se destaca por integrar herramientas de IA en todas sus carreras profesionales desde el primer ciclo.
2. Universidad de San Martín de Porres (USMP): Inauguró el primer laboratorio de inteligencia artificial en el Perú en colaboración con Huawei

(IALAP). El laboratorio utiliza la plataforma ModelArts para la investigación en analítica de datos, big data y gestión de datos.

3. Pontificia Universidad Católica del Perú (PUCP): Imparte diplomados de seis meses en Inteligencia Artificial, con un enfoque que combina fundamentos teóricos y aplicaciones industriales reales.
4. Universidad del Pacífico (UP): Ofrece programas de especialización en Inteligencia Artificial aplicada al Sector Público a través de su Escuela de Gestión Pública, con enfoque en la modernización del Estado.

Recursos bibliográficos y bibliotecas técnicas

Para el estudio autodidacta y la investigación académica, Lima cuenta con colecciones bibliográficas extensas, muchas de ellas con acceso digital.

- Biblioteca Nacional del Perú (BNP): Ofrece un catálogo general en línea que incluye libros técnicos, hemerografía y recursos electrónicos. Su plataforma BNP Digital permite acceder a miles de documentos digitalizados para estudiantes y profesionales.
- Biblioteca de la Escuela Nacional de Administración Pública (ENAP): Ubicada en Jesús María (Av. Cuba 699), ofrece servicios de búsqueda especializada y capacitaciones para fortalecer las competencias en el sector público.
- Biblioteca de la Universidad del Pacífico: Cuenta con guías sobre producción académica en inteligencia artificial y recursos para la investigación
- Biblioteca del Organismo de Evaluación y Fiscalización Ambiental (OEFA): Ubicada en Jesús María, dispone de recursos bibliográficos y organiza programas

sobre el uso de herramientas de IA en la investigación científica.

Librerías especializadas en textos técnicos

La adquisición de libros fundamentales, como AIMA o el texto de Deep Learning de Goodfellow, puede realizarse en librerías con secciones técnicas dedicadas.

- Librería SBS (Special Book Services): Tiene sedes en la Universidad del Pacífico (Jesús María) y en San Marcos. Su catálogo incluye libros de ingeniería de software, de matemáticas para ingeniería e inteligencia artificial.
- Librería Communitas: Ubicada cerca de Jesús María (Av. Dos de Mayo 1690, San Isidro), cuenta con un amplio rubro en ciencia y tecnología, además de literatura técnica internacional.
- Librerías Crisol: Con sedes en centros comerciales como Real Plaza Salaverry (Jesús María), ofrece una sección de libros técnicos y profesionales.

Ética, seguridad y el futuro de la inteligencia artificial

A medida que la inteligencia artificial se integra más profundamente en la sociedad, los temas de seguridad y ética se vuelven inseparables del desarrollo técnico. La cuarta edición de AIMA dedica capítulos enteros a la filosofía y a la seguridad de la IA, abordando el impacto del desempleo tecnológico y los riesgos de los sistemas superinteligentes.

Uno de los mayores desafíos es garantizar que los sistemas de IA sean justos y confiables. Esto implica no solo eliminar los sesgos en los datos de

entrenamiento, sino también asegurar que los objetivos del sistema coincidan con los valores humanos. La IA neuro-simbólica se perfila aquí como una solución clave, ya que permite imponer restricciones éticas explícitas mediante reglas simbólicas que el sistema no puede ignorar, independientemente de los patrones que aprenda a partir de los datos (Smuha, 2025).

La evolución desde la IA basada en reglas hacia el aprendizaje profundo y, ahora, hacia la síntesis neuro-simbólica refleja un intento de capturar la complejidad de la inteligencia humana. El futuro del campo reside en desarrollar agentes capaces de percibir con la agudeza del aprendizaje profundo, razonar con la precisión de la lógica formal y operar en entornos desconocidos con la adaptabilidad del aprendizaje por refuerzo.

La integración de estos enfoques no solo promete sistemas más potentes, sino también más seguros y transparentes, capaces de colaborar de manera efectiva con los seres humanos en la resolución de los problemas más complejos de nuestra era. En este camino, la educación continua y el acceso a recursos técnicos, como los disponibles en los nodos académicos de Lima, serán fundamentales para asegurar un desarrollo tecnológico equitativo y responsable.

A pesar del auge del aprendizaje profundo, los algoritmos clásicos de búsqueda siguen siendo la base de la planificación y el razonamiento lógico en sistemas de IA. Estos se dividen principalmente en búsquedas no informadas e informadas.

Los métodos de búsqueda no informada, como la búsqueda en anchura

(BFS) y la búsqueda en profundidad (DFS), exploran el espacio de estados sin conocer la proximidad de la meta. El algoritmo de Dijkstra, o búsqueda de costo uniforme, expande el nodo con el menor costo acumulado, garantizando la optimalidad si los costos son positivos (Escobar & Giraldo, 2005).

La búsqueda informada utiliza funciones heurísticas para guiar el proceso. El algoritmo A* es el estándar en este ámbito, combinando el costo acumulado $g(n)$ con una estimación del costo restante $h(n)$ para encontrar el camino óptimo de manera eficiente. En entornos adversarios, como los juegos de suma cero, se emplean el algoritmo Minimax y la poda Alfa-Beta para reducir drásticamente el espacio de búsqueda al descartar ramas que no afectan la decisión final.

1.4 El paradigma del aprendizaje profundo (Deep Learning)

El aprendizaje profundo ha transformado la capacidad de las máquinas para procesar datos no estructurados como imágenes, audio y texto. Se define como una forma de aprendizaje automático que permite a las computadoras aprender de la experiencia a través de una jerarquía de conceptos.

El éxito de las redes neuronales profundas se basa en tres áreas matemáticas críticas: el álgebra lineal, la probabilidad y el cálculo multivariable. Las redes neuronales convolucionales (CNN) dominan el procesamiento visual mediante capas que detectan características espaciales, mientras que las redes neuronales recurrentes (RNN) y sus variantes, como LSTM, fueron el estándar para datos secuenciales hasta la llegada de los Transformers.

La arquitectura Transformer, introducida en 2017, revolucionó el campo al eliminar la necesidad de procesamiento secuencial, permitiendo el paralelismo masivo y capturando dependencias de largo alcance mediante el mecanismo de atención. Esta innovación es la base de los modelos de lenguaje a gran escala (LLM) actuales.

Procesamiento de Lenguaje Natural (NLP) y Redes Generativas (GAN)

La convergencia entre el aprendizaje profundo y el procesamiento de señales ha dado lugar a una nueva era en la interacción humano-máquina y en la creación de contenido sintético.

Evolución del NLP: de reglas a modelos de representación

Hoy en día, el campo ha transitado de una disciplina impulsada por características manuales a otra centrada en la representación, donde modelos como BERT (Bidirectional Encoder Representations from Transformers) y GPT (Generative Pre-trained Transformer) capturan semántica, tono y contexto con una precisión sin precedentes. Hoy, las tendencias de investigación en conferencias como ACL destacan un auge del trabajo multimodal (visión-lenguaje), donde la percepción clásica se replantea como una tarea de seguimiento de instrucciones y de razonamiento en múltiples pasos.

Redes Generativas Antagónicas (GAN): El juego de la creación

Introducidas por Ian Goodfellow en 2014, las GAN representan uno de los avances más influyentes en el campo de los modelos generativos. Su arquitectura consiste en dos redes neuronales que compiten en un juego de suma cero de tipo minimax:

1. **Generador (G):** Intenta generar muestras de datos (imágenes, texto, audio) que imiten la distribución de los datos reales para engañar al discriminador.

2. **Discriminador (D):** Evalúa si una muestra proviene de los datos de entrenamiento reales o del generador.

El proceso matemático de entrenamiento se define por la función de valor:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{data}(x)} + \mathbb{E}_{z \sim p_z(z)}$$

En síntesis, la investigación en GAN se ha consolidado en torno a la controlabilidad, la síntesis de video y la generación de imágenes de ultra alta resolución (4K), superando los desafíos históricos de inestabilidad en el entrenamiento y de colapso de modo.

A pesar de los avances en el aprendizaje profundo, los modelos neuronales puros a menudo carecen de interpretabilidad y sufren de opacidad. La inteligencia artificial neuro-simbólica surge como un enfoque híbrido que combina la robustez perceptiva del aprendizaje profundo con la capacidad de razonamiento lógico de la IA simbólica clásica.

Sistemas como AlphaGeometry ejemplifican este poder al combinar un modelo de lenguaje neuronal (intuición) con un motor de deducción simbólica (lógica). Asimismo, las Logic Tensor Networks (LTN) permiten traducir las fórmulas de lógica de primer orden a términos diferenciables, asegurando que

el sistema aprenda patrones de datos mientras respeta restricciones lógicas absolutas (Wang et al., 2025).

Capítulo II
Inteligencia Artificial: Explorando Enfoques Modernos y el Futuro del Aprendizaje Profundo

La historia de la ciencia de datos y la inteligencia artificial (IA) es un testimonio de la aceleración tecnológica exponencial. En sus albores, durante la década de 1950, las predicciones de Alan Turing sobre las máquinas pensantes sentaron las bases filosóficas y teóricas que guiarían a generaciones de investigadores. Posteriormente, durante las décadas de 1960 y 1970, la sociedad dependía de análisis manuales rudimentarios, con un volumen de datos tan escaso que podía procesarse por equipos humanos minúsculos mediante los primeros programas estadísticos. No fue sino hasta la década de 1990 que el término ciencia de datos se consolidó formalmente, marcando la transición hacia la era digital y la gestión de bases de datos masivas.

El verdadero punto de inflexión arquitectónico se gestó gracias a los pioneros de las redes neuronales, como Geoffrey Hinton y Yann LeCun, quienes, entre las décadas de 1980 y 2000, pavimentaron el camino para los modelos generativos modernos. Esta labor fundacional impulsó el auge del aprendizaje profundo (Deep Learning) en la década de 2010, desencadenando avances monumentales en el procesamiento del lenguaje natural (NLP), la generación de imágenes y los diagnósticos médicos mediante la segmentación de alta precisión. Sin embargo, la trayectoria del aprendizaje profundo ha alcanzado un nuevo umbral crítico. Tras años de expansión desenfrenada y apuestas corporativas multimillonarias, las proyecciones a mediano plazo coinciden en que la industria está abandonando la era del evangelismo de la IA para adentrarse en la de la evaluación rigurosa de la IA.

En la actualidad, las interrogantes científicas ya no giran exclusivamente en torno a la viabilidad teórica de las tareas cognitivas computacionales, sino que se centran en evaluar con precisión cómo operan estos sistemas, a qué costo energético y bajo qué marcos de equidad y responsabilidad. El ecosistema de la inteligencia artificial exige ahora modelos más pequeños, altamente eficientes, dotados de capacidades profundas de razonamiento causal y de nuevas formas de inteligencia encarnada (embodied intelligence). A medida que los requisitos de hardware de los Modelos de Lenguaje Grande (LLMs) se aproximan a los límites de la infraestructura computacional convencional, la innovación se ve obligada a diversificarse hacia nuevas arquitecturas computacionales, como el modelado de mundos, la computación neuromórfica y la convergencia neuro-simbólica.

El presente capítulo proporciona una disección exhaustiva del estado del arte en inteligencia artificial, analizando meticulosamente los enfoques algorítmicos modernos, los desafíos de sostenibilidad, los dilemas de la interpretabilidad mecanicista y la reconfiguración del panorama regulatorio global que definirá el camino hacia la inteligencia artificial general (AGI).

2.1 Modelos Fundacionales, Aprendizaje Autosupervisado y Evolución Arquitectónica

El núcleo del progreso reciente en inteligencia artificial reside en la consolidación de los Modelos Fundacionales (FMs), estructuras computacionales masivas impulsadas por el Aprendizaje Autosupervisado (Self-Supervised Learning o SSL). Históricamente, los métodos tradicionales de aprendizaje automático exhibían limitaciones paralizantes: dependían de la extracción manual de características (handcrafted features), eran altamente sensibles a las variaciones en la calidad de los datos según el entorno de grabación y mostraban una capacidad deficiente para modelar la naturaleza no lineal y dinámica de fenómenos complejos (Voeneky et al.,

2022).

Transición hacia Dominios Biológicos y Médicos

El aprendizaje autosupervisado ha transformado esta dinámica al permitir que las redes neuronales aprendan representaciones latentes y significativas directamente a partir de vastos corpus de datos no etiquetados, mitigando el cuello de botella que representa la anotación humana. Tras redefinir la visión por computadora y el procesamiento del lenguaje, la comunidad científica ha aplicado estas técnicas fundacionales a dominios de extrema complejidad, como la imagenología médica y el análisis de señales cerebrales (Taherdoost, 2024).

En el ámbito neurológico, por ejemplo, los modelos enfrentan desafíos únicos caracterizados por niveles de ruido exorbitantes, relaciones señal-ruido marginales y una variabilidad intersujetos que inhabilitan a los algoritmos supervisados clásicos. El SSL ofrece una solución elegante para desarrollar modelos fundacionales específicos para el cerebro, integrando señales neuronales con otras modalidades en marcos multimodales para tareas que abarcan desde la imaginería motora hasta los potenciales evocados visuales y la carga cognitiva. Paralelamente, en el ámbito quirúrgico, los esfuerzos de investigación actuales incluyen estudios exploratorios a gran escala sobre métodos SSL aplicados a la visión por computadora quirúrgica, preentrenando modelos en conjuntos de datos públicos y privados extensos para evaluar el impacto estructural de la composición de los datos.

Eficiencia Paramétrica y Aprendizaje Continuo

A pesar de la primacía del aprendizaje autosupervisado, la optimización computacional ha revelado matices empíricos inesperados. Estudios rigurosos recientes han demostrado que el ajuste fino eficiente en parámetros (Parameter-Efficient Fine-Tuning o PEFT), ejecutado exclusivamente sobre datos etiquetados, a

menudo iguala el rendimiento algorítmico del SSL sin necesidad de procesar los inmensos volúmenes de datos no etiquetados (Watson et al., 2024). Esta observación ha motivado a los investigadores a revisitar el autoentrenamiento (self-training), una línea conceptual de base en la que un modelo PEFT supervisado se utiliza para generar pseudoetiquetas en datos no estructurados, lo que permite un aprendizaje iterativo de alta eficiencia.

Asimismo, las técnicas de autosupervisión han demostrado ser vitales para resolver el problema de la adaptación continua (Continual Learning, CL). El Entrenamiento Previo Continuo (CPT) depende en gran medida de objetivos de aprendizaje como la predicción de tokens enmascarados, la predicción de la siguiente oración y el aprendizaje contrastivo imagen-texto. La evidencia empírica documentada subraya que continuar preentrenando grandes modelos de manera autosupervisada sobre flujos de datos en tiempo real produce una retención de conocimiento estructuralmente superior a la del entrenamiento supervisado, mitigando el olvido catastrófico y facilitando la reutilización de los modelos sin requerir reentrenamientos desde cero.

La Evolución de la Arquitectura del Transformador

Para comprender la magnitud de estos avances, es imperativo analizar la evolución mecánica de las arquitecturas subyacentes. Aunque la arquitectura del Transformador, introducida en 2017, sigue siendo el cimiento de la IA generativa hoy en día, sus componentes internos han sufrido mutaciones drásticas orientadas a la eficiencia y a la capacidad de generalización (ver Tabla 3).

Tabla 3: Evolución mecánica de la inteligencia artificial generativa

Componente Arquitectónico	Modelos Tempranos (ej. GPT-2)	Modelos Modernos (Estado del Arte 2025)	Justificación de la Evolución
Mecanismo de Atención	Atención Multi-Cabezal (MHA)	Atención de Consulta Agrupada (GQA) y Ventana Deslizante	Reducción drástica del uso de memoria y aceleración de la inferencia para *prompts* extensos.
Incrustaciones Posicionales	Posicionales Absolutas (Aprendidas)	Incrustaciones de posición rotatoria (RoPE)	Mejora sustancial en la generalización y en la comprensión de contextos extremadamente extensos.
Estructura Paramétrica	Arquitectura Densa (Activación total)	Mezcla de Expertos (MoE) Dispersa	Aumento masivo de la capacidad de conocimiento, activando únicamente los módulos especializados mediante token.
Capas Feed-Forward	Función de activación GELU	Unidades de activación compuerta (SwiGLU)	Mayor expresividad matemática y un rendimiento superior gracias a un recuento de parámetros más eficiente.
Normalización	LayerNorm	RMSNorm (Root Mean Square Normalization)	Simplificación computacional que facilita una ejecución exponencialmente más rápida en GPUs.

2.2 Interpretabilidad Mecanicista y el Desafío de la Transparencia Algorítmica

A medida que las arquitecturas se vuelven infinitamente más complejas, surge una paradoja epistemológica fundamental: la comunidad científica no comprende cabalmente cómo funcionan internamente los modelos de lenguaje masivos. Este problema de la caja negra ha dado origen al campo de la interpretabilidad mecanicista (Mechanistic Interpretability, o MI), una disciplina que busca aplicar la ingeniería inversa a las redes neuronales para mapear cómo el razonamiento emerge a partir de interacciones matemáticas elementales.

Organizaciones líderes en seguridad algorítmica dedican recursos masivos a descifrar los circuitos transformadores para garantizar un comportamiento confiable y alineado. Las investigaciones de vanguardia en 2025 han documentado hallazgos extraordinarios sobre la conciencia introspectiva emergente en los LLMs, lo que sugiere que estos sistemas desarrollan la capacidad de introspeccionar sus propios estados internos.

Entre los avances metodológicos más destacados se encuentra el desarrollo de Oráculos de Activación, sistemas mediante los cuales se entrena a los modelos de lenguaje para que respondan preguntas sobre sus propias activaciones neuronales en lenguaje natural, lo que proporciona una ventana interpretativa sobre su razonamiento latente. Asimismo, la investigación sobre la geometría de tareas cognitivas ha demostrado que cuando los modelos ejecutan tareas fundamentales como el conteo, manipulan estructuras geométricas complejas (variedades topológicas o *manifolds*) subyacentes a sus mecanismos de atención.

El análisis de la computación de la atención mediante la interacción entre características (features) y la ponderación de la interferencia ha llevado al diseño de Mezclas Dispersas de Transformadas Lineales (MOLT), un nuevo enfoque para transcodificadores que busca desentrañar el problema de la superposición, en el que las redes neuronales agrupan múltiples conceptos en un número limitado de

dimensiones (Sprockel et al., 2025). Este mapeo detallado de circuitos dispersos es vital no solo para la auditoría de alineación automatizada, sino también para evitar que arquitecturas no reguladas generen fallos sistémicos indetectables.

2.3 Más Allá del Modelado Autorregresivo: Modelos de Mundo y Arquitecturas JEPA

El dogma imperante en el aprendizaje profundo reciente ha sido el de las Leyes de Escalado (Scaling Laws): la suposición de que incrementar los parámetros del modelo, el tamaño del conjunto de datos y la potencia de cálculo conduce ineludiblemente a mejoras proporcionales en el rendimiento, un fenómeno documentado desde la robótica hasta los agentes que juegan videojuegos mediante aprendizaje por imitación y modelado de entornos (Porcelli, 2020). Sin embargo, investigaciones críticas de 2025 revelan que la intuición de que lo más grande es siempre mejor es una simplificación peligrosa.

Las limitaciones de las leyes de escalado

Si bien es cierto que las leyes de potencias se manifiestan en la caída de la pérdida (loss) en función del tamaño óptimo del modelo, los coeficientes y las pendientes de estas curvas de escalabilidad están fuertemente determinados por factores microscópicos: la metodología de tokenización, la naturaleza específica de la tarea y la arquitectura de la red. Además, la evaluación del rendimiento basada en mediciones de un solo intento (one-off accuracies) resulta profundamente engañosa.

A nivel fundamental, existe un límite teórico a la eficiencia con la que los modelos autorregresivos pueden aprender conceptos. Al depender exclusivamente de la predicción iterativa del siguiente token, los LLMs carecen de una representación genuina de la física, del sentido común y de capacidades de planificación a largo plazo.

El Paradigma V-JEPA y la Inteligencia de Máquina Autónoma

En respuesta a estas barreras teóricas, el campo de la inteligencia artificial está virando hacia el desarrollo de Modelos de Mundo (World Models), un área impulsada con fuerza por pioneros como Yann LeCun, quien recientemente fundó Advanced Machine Intelligence (AMI) Labs—una empresa que apunta a valoraciones multimillonarias y está centrada en la comercialización de esta tecnología predictiva (Voeneky et al., 2022).

El núcleo técnico de esta revolución es la Arquitectura Predictiva de Incrustación Conjunta (JEPA), introducida conceptualmente en 2022 y expandida en los años posteriores a dominios multimodales como VL-JEPA (Visión-Lenguaje) y V-JEPA (Video). A diferencia de los modelos generativos tradicionales que malgastan recursos inmensos intentando reconstruir cada píxel faltante en una imagen o cada ruido estocástico en un entorno de video, las arquitecturas JEPA aprenden a predecir representaciones abstractas (latentes) de la realidad.

Este enfoque dota al modelo de la capacidad de ignorar detalles irrelevantes o caóticos del entorno y de enfocarse en las dinámicas causales fundamentales. La integración de la teoría matemática subyacente (LeJEPA) ha fortalecido este marco, estableciendo los Modelos de Mundo como el puente indispensable para la transición tecnológica desde el procesamiento de texto estático hasta la robótica avanzada, la inteligencia artificial encarnada y, en última instancia, la autonomía cognitiva.

2.4 Inteligencia Artificial Causal: Inferencia, Razonamiento y Aplicaciones Prácticas

El aprendizaje profundo tradicional ha triunfado explotando patrones estadísticos masivos; no obstante, su ceguera inherente ante la causalidad —su incapacidad para distinguir entre una correlación espuria y una relación de causa-

efecto— compromete severamente su confiabilidad en sistemas críticos (Russel y Norvig, 2004). Como antídoto, la Inteligencia Artificial Causal (Causal AI) ha escalado aceleradamente, reconociéndose en el Ciclo de Expectativas de Gartner y destacando en encuestas a líderes de la industria como la técnica emergente prioritaria para la adopción corporativa y el despliegue masivo.

Técnicas Subyacentes y Evolución Metodológica

La IA causal representa una simbiosis metodológica que fusiona las redes neuronales profundas con los axiomas de la inferencia estadística para producir sistemas robustos, justos e interpretables. La integración con Grandes Modelos de Lenguaje (LLMs) y el avance en el descubrimiento causal en tiempo real facilitan la resolución del prolongado problema de la caja negra y la reducción sistémica de sesgos (ver Tabla 4).

Tabla 4: Mecanismo, aplicación y propósito de técnicas de inferencia causal

Técnica de Inferencia Causal	Mecanismo y Aplicación Analítica	Propósito en el Modelado Algorítmico
Ensayos controlados aleatorizados (RCTs)	Asignación aleatoria de sujetos a los grupos de tratamiento y de control.	Considerado el estándar de oro para aislar intervenciones causales puras.
Razonamiento Contrafactual	Evaluación matemática y simulación de escenarios alternativos (¿Qué habría pasado si…?).	Estimación de dimensiones no observadas y evaluación prospectiva de políticas o tratamientos.
Emparejamiento por Puntaje de Propensión	Construcción algorítmica de grupos de control artificiales en bases de datos masivas.	Control riguroso de las variables de confusión en estudios de datos meramente observacionales.

Variables Instrumentales y Discontinuidad	Diseño de regresiones estructuradas en torno a intervenciones externas.	Estimación de causalidad en entornos en los que la experimentación controlada resulta logística o éticamente inviable.

Disrupción en la salud y sistemas dinámicos

En los sectores de la atención médica y de las ciencias biológicas, la IA causal está impulsando la medicina personalizada. Los sistemas causales analizan historias clínicas para simular contrafactuales y prever matemáticamente cómo fluctuaría la salud de un paciente ante terapias alternativas. El impacto clínico es contundente: investigaciones recientes validaron que algoritmos diagnósticos fundamentados en principios causales alcanzaron un rendimiento equivalente al de expertos humanos, ubicándose en el 25% superior de los facultativos médicos en precisión diagnóstica, superando ampliamente a los modelos de aprendizaje automático convencionales.

Expertos visionarios de instituciones de salud proyectan que la IA, mediante simulaciones fluidas y microfluídica, diagnosticará la ceguera futura por retinopatía diabética analizando la dinámica celular en la sangre, así como acelerará el diagnóstico temprano de trastornos hereditarios como la anemia falciforme y la eliptocitosis mediante la inferencia estructural de la morfología de los glóbulos rojos. Adicionalmente, las instituciones hospitalarias despliegan IA causal para discernir las causas raíz de las readmisiones y reorientar los presupuestos de camas hospitalarias hacia la atención de seguimiento posalta, fundamentada en datos empíricos (Enriquez & Raraz, 2024).

Más allá de la medicina, la inferencia causal ha demostrado superioridad en entornos altamente dinámicos y caóticos, como las simulaciones de videojuegos de estrategia en tiempo real (RTS). Investigaciones que utilizaron la API de StarCraft II, el modelado mediante grafos causales (PyMC) y redes neuronales profundas

(PyTorch) resolvieron problemas de escasez de datos mediante redes generativas antagonistas tabulares condicionales (CGAN). Los sistemas híbridos que inyectaron conocimientos causales probabilísticos sobre tácticas militares e interacciones de recursos directamente en las redes neuronales produjeron predicciones un 1,1 % más precisas y adaptativas que los enfoques puramente profundos, mitigando los sesgos ofensivos intrínsecos de los datos de entrenamiento base.

2.5 Alineación, Sesgos Algorítmicos y la Paradoja de los Datos Sintéticos

El método predominante para forjar el comportamiento deseable en los LLMs es el Aprendizaje por Refuerzo a partir de Retroalimentación Humana (RLHF). Sin embargo, análisis matemáticos profundos revelan que este esquema padece un sesgo algorítmico inherente, derivado del uso de la divergencia de Kullback-Leibler (KL) como mecanismo de regularización en el proceso de optimización. Esta estructura matemática propicia un fenómeno peligroso denominado colapso de preferencias, en el que las inclinaciones minoritarias, sensibilidades culturales divergentes o voces marginadas son virtualmente asimiladas e ignoradas por la política del modelo, lo que impulsa una homogeneización artificial hacia la media mayoritaria.

La diversidad inmanente a las preferencias humanas, moldeada por la subjetividad del contexto, las ambigüedades lingüísticas y el bagaje sociodemográfico, queda borrada en este esquema. Además, la naturaleza algorítmica de la selección de código abierto se encuentra desproporcionadamente sesgada hacia conjuntos de datos occidentales y en inglés, marginando estructuralmente al Sur Global. Para corregir esta falla sistémica, la frontera de la investigación propone el RLHF de Emparejamiento de Preferencias (Preference Matching RLHF, o PM-RLHF), un enfoque matemático basado en los modelos de Plackett-Luce y Bradley-Terry-Luce.

El PM-RLHF utiliza el logaritmo negativo de la distribución de probabilidad de la política del LLM sobre las respuestas, resuelto mediante ecuaciones diferenciales ordinarias, para equilibrar explícitamente la maximización de la recompensa con la diversificación de las salidas. Evaluaciones empíricas sobre las arquitecturas de la familia Llama y OPT han confirmado que la integración condicional de PM-RLHF eleva el nivel de alineación con la diversidad humana real entre un 29% y un 41% frente a las métricas estándar.

La Trampa de la Confianza Sintética

En el intento de eludir los sesgos estructurales de los datos del mundo real, proteger la privacidad y sortear la inminente escasez de corpora lingüísticos disponibles en internet, la comunidad global de desarrollo de IA ha recurrido de manera agresiva al uso de datos sintéticos. La capacidad de crear réplicas matemáticas que preservan las propiedades estadísticas de las distribuciones originales sin comprometer los puntos de datos individuales permite diseñar intervenciones justas. Los datos sintéticos, fundamentados en restricciones de equidad y de ponderación causal, tienen el potencial de eliminar covariables discriminatorias y democratizar el acceso al entrenamiento de algoritmos en entornos de alto riesgo.

No obstante, esta panacea estadística entraña riesgos sistémicos agudos documentados a lo largo del año 2025. El peligro primario es la emergencia de la Confianza Sintética (Synthetic Trust): una seguridad injustificada depositada en modelos entrenados con datos generados algorítmicamente que fracasan drásticamente en preservar las realidades demográficas o la validez empírica.

El entrenamiento endogámico prolongado —donde los modelos ingieren continuamente las salidas de otros modelos generativos (como Redes Generativas Antagónicas o modelos de difusión)— precipita una rápida degradación topológica

del espacio latente, erosionando la capacidad del sistema para reconocer escenarios novedosos del mundo real e introduciendo nuevos sesgos impredecibles. Los investigadores exigen la implementación urgente de salvaguardas arquitectónicas: la integración inquebrantable de auditorías de fragilidad, la obligatoriedad de revelar la procedencia sintética y el uso forzoso de enfoques híbridos provistos de mecanismos de autocorrección arraigados en datos orgánicos puros.

2.6 La Frontera Neuro-Simbólica y la Revolución de la Computación Neuromórfica

Si el aprendizaje profundo estándar y la inferencia causal representan los peldaños actuales de la IA, la academia define el horizonte inmediato como el tercer verano de la IA, una época marcada por la sinergia entre la Inteligencia Artificial Neuro-Simbólica y el hardware biológicamente inspirado (Liang et al., 2025).

La Fusión de los Sistemas Cognitivos 1 y 2

Desde una perspectiva cognitiva, las arquitecturas puramente neuronales emulan eficientemente el Sistema 1 del pensamiento humano: procesos rápidos, paralelos e intuitivos enfocados en el reconocimiento de patrones. Sin embargo, son inherentemente frágiles ante tareas de lógica abstracta que requieren la manipulación secuencial de símbolos, un dominio gobernado por el Sistema 2 (lento, deliberado y lógico).

La IA neurosimbólica persigue la integración de ambos paradigmas, inyectando conocimiento previo lógico explícito en la plasticidad diferencial de las redes subsimbólicas. Una revisión sistemática exhaustiva de los últimos años, que analizó cientos de trabajos revisados por pares, cuantificó empíricamente la distribución del esfuerzo investigativo para definir el estado del arte de esta frontera (ver Tabla 5).

Tabla 5: Distribución del esfuerzo investigativo por áreas tipo

Área de Investigación Fundamental	Esfuerzo Representado	Descripción del Foco de Estudio
Aprendizaje e Inferencia	63%	Combinación de aprendizaje profundo diferencial con métodos deductivos dinámicos y multifuente.
Representación del Conocimiento	44%	Integración de incrustaciones neuronales continuas con grafos de conocimiento explícitos y ontologías estructuradas.
Lógica y Razonamiento	35%	Incorporación de la semántica lógica y probabilística como funciones de pérdida o como limitadores del espacio latente en redes.
Explicabilidad y Confianza	28%	Desarrollo de la trazabilidad en los procesos neuronales, clasificado como un vacío notable y crítico para un despliegue confiable.
Meta-Cognición	5%	El eslabón perdido: dotar al agente de la capacidad de evaluar, auditar y reestructurar sus propios razonamientos en tiempo real.

La incipiente exploración de la metacognición demuestra empíricamente que, a pesar de la euforia corporativa, las aproximaciones actuales a la cognición humana

siguen siendo mecánicamente simplistas. Lograr la autonomía en entornos volátiles exigirá sistemas metacognitivos adaptativos que entrelacen arquitecturas cognitivas complejas con LLMs para facilitar la resolución introspectiva de conflictos lógicos (Walker et al., 2025).

El Cuello de Botella Energético y la Computación Neuromórfica

La ambición de escalar estas arquitecturas choca frontalmente con la segunda ley de la termodinámica. La inteligencia artificial actual basada en máquinas de von Neumann es una consumidora de energía insostenible. Modelos predictivos del Departamento de Energía de Estados Unidos (DOE) estiman proyecciones escalofriantes: el costo operativo y energético global para alimentar hipermodelos LLM podría superar el Producto Interno Bruto total de Estados Unidos para el año 2027, con facturas eléctricas anuales teóricas que alcanzarían los billones de dólares. Aunque estas estimaciones pueden representar escenarios límite, evidencian que el escalado mediante fuerza bruta basada en transistores binarios masivos resulta ecológicamente inviable.

El escape de este callejón físico es la computación neuromórfica. En oposición radical al procesamiento binario convencional, los procesadores neuromórficos imitan el isomorfismo estructural del cerebro biológico, ejecutando redes fotónicas y eléctricas asíncronas de bajísimo consumo de energía. Mientras un centro de datos corporativo requiere decenas de megavatios para entrenar millones de parámetros, los neurocientíficos computacionales de laboratorios nacionales como Los Álamos proyectan hardware neuromórfico capaz de ejecutar inferencias cognitivas avanzadas consumiendo tan solo 20 vatios —el equivalente calórico diario del cerebro humano o la electricidad que consumen dos pequeñas bombillas LED—.

Al superar la dicotomía de memoria-procesador de Von Neumann, estos

procesadores espaciales logran reducciones drásticas en los requerimientos de cálculo de entrenamiento y permiten el aprendizaje y la inferencia in situ en dispositivos de borde (mobile edge computing). Esta tecnología es indispensable para la realización de robótica encarnada eficiente, posibilitando, en palabras de la investigación estatal, el diseño de agentes de drones del tamaño de un mosquito equipados con la adaptabilidad plástica natural equivalente a la de su homólogo biológico.

2.7 La Crisis de Sostenibilidad y la Eficiencia Algorítmica (Compresión y Poda)

El frenesí por implementar capacidades cognitivas artificiales ha incubado una crisis ambiental sin precedentes, que el Foro Económico Mundial tipifica como la Paradoja Energética de la Inteligencia Artificial. A la par que la IA promete avances en herramientas de descarbonización y en el modelado climático científico, la huella ecológica generada por el despliegue físico de la misma infraestructura computacional (GPUs, SSDs, centros de datos masivos) es descomunal.

Radiografía del Impacto Ambiental (Emisiones e Hídrico)

La infraestructura subyacente a los modelos de miles de millones de parámetros exhibe un metabolismo industrial destructivo. La huella abarca desde la minería contaminante requerida para fabricar obleas de silicio complejas y la síntesis química tóxica hasta el asombroso drenaje de las redes eléctricas continentales (Amzil et al., 2025). A modo de ilustración histórica, el mero proceso de preentrenamiento del modelo GPT-3 consumió alrededor de 1,287 megavatios-hora (MWh) —suficiente para abastecer 120 hogares estándar durante un año— y emitió aproximadamente 552 toneladas de dióxido de carbono a la atmósfera. Durante la etapa operativa (inferencia), la disparidad es dramática: procesar una simple consulta mediante un motor generativo agota cinco veces más recursos eléctricos que una búsqueda indexada convencional.

Las métricas macroeconómicas de la FP Analytics y organismos análogos muestran que la capacidad global de consumo de los centros de datos, situada entre 415 y 460 teravatios-hora (TWh) en 2022 y 2024, se catapultará exponencialmente hasta umbrales entre 1.200 y 1.700 TWh para el año 2035; una carga electrotérmica equiparable a la de la economía de la India en su totalidad (Silva et al., 2025). Exacerbando la crisis climática, el hardware generativo demanda densidades de potencia hasta ocho veces superiores a las de la computación tradicional, lo que impulsa un estrés masivo sobre los suministros hídricos. El enfriamiento líquido avanzado de servidores evapora, en promedio, dos litros de agua dulce municipal por cada kilovatio-hora de energía procesada, lo que altera irreparablemente los ecosistemas perimetrales y propicia la interrupción de los suministros locales.

Para mitigar la sangría, institutos de investigación climática desarrollan tácticas operativas que intercalan la IA con la matriz energética; el estrangulamiento de potencia (power capping) sobre las GPUs restringe el sobrecalentamiento al recortar el consumo eléctrico en un 20% a 30%, con penalizaciones de rendimiento mínimas, mientras que rutinas de balanceo temporal derivan cargas computacionales masivas hacia franjas horarias de baja intensidad de carbono, reduciendo las emisiones directas en un 80%.

El Estado del Arte en Compresión y Cuantización

La barrera energética y la imperiosa necesidad de desplegar arquitecturas en el borde (edge) impulsan un ecosistema de investigación hiperactivo centrado en técnicas de compresión de modelos matemáticos que reduzcan la huella de memoria sin mutilar el desempeño heurístico (ver Tabla 6).

Tabla 6: Técnicas de compresión de modelos matemáticos

Metodología de Compresión	Principio Algorítmico y Evolución Técnica	Impacto Estructural Reciente
Cuantización de Precisión Mixta (AWQ)	Traducción matemática de la continuidad de punto flotante de alta resolución (FP32) a formatos numéricos enteros discretos reducidos.	La Cuantización Consciente de la Activación aísla el 1% de pesos hipercríticos manteniéndolos puros, mientras colapsa el 99% restante a resoluciones de 4 bits, logrando una compresión de 8x sin atrofiar LLMs.
Cuantización extrema a 1 bit	Explotación de parámetros geométricos compartidos, donde las matrices de pesos se discretizan exclusivamente en escalares de 1 y -1.	Tras reconstruir los parámetros originales de inferencia a través de productos punto, destruye las cargas de VRAM hasta en un 90%, lo que abre la frontera de LLMs viables en teléfonos.
Poda No Estructurada (Unstructured Pruning)	Erradicación asimétrica de conexiones neuronales individuales basada en algoritmos iterativos desarrollados en los años 80 y 90 (OBD/OBS).	Crea arquitecturas de dispersión geométrica irregular (hipótesis del boleto de la Lotería). Preserva la precisión crítica, pero exige librerías operativas especializadas que fracturan el hardware de aceleración.
Poda Estructurada (Structured Pruning)	Supresión sistemática de filtros macroscópicos, canales volumétricos o módulos de atención completos.	Transige fraccionariamente con la precisión general en pro de crear matrices estructuralmente densas, lo que permite a CPU/GPU ejecutar operaciones de aritmética con inmensa celeridad y fluidez.

El pináculo investigativo actual promueve una simbiosis definitiva: el codiseño de software-hardware. Modelos híbridos modernos como ACDNet y MobileNet-v3-small fusionan algoritmos de poda secuencial con cuantizaciones severas (8 bits enteros o menores), incrustados directamente en las arquitecturas lógicas de las unidades de silicio periféricas, solventando cuellos de botella de hardware mediante operaciones de desplazamiento de bits a micropotencia de bajo voltaje.

2.8 Gobernanza Global y Soberanía de la Inteligencia Artificial (2025-2027)

Lejos de circunscribirse al ámbito académico, las implicaciones económicas y existenciales del control cognitivo de la inteligencia artificial han suscitado una fractura geopolítica sin precedentes en la historia. De acuerdo con reportes del Foro Privado de Datos, alrededor de mil legislaciones se introdujeron a nivel global en 2025. Entre 2025 y 2027, las macrorregiones económicas mundiales adoptarán doctrinas contrapuestas sobre el grado óptimo de supervisión regulatoria y de censura algorítmica.

La Arquitectura Regulatoria Previsora de la Unión Europea

La Unión Europea ostenta el marco normativo más sofisticado del mundo: la Ley de Inteligencia Artificial de la UE (EU AI Act). Su mecanismo se basa en una categorización piramidal del riesgo. A partir de febrero de 2025, la ley impuso prohibiciones inapelables sobre prácticas que transgreden los derechos humanos fundamentales; sistemas de puntuación ciudadana, el despliegue de análisis emocional biométrico en centros laborales o inferencias sociodemográficas indiscriminadas en tiempo real quedaron totalmente proscritos en el bloque europeo (Aparicio, 2025).

Para agosto de 2025, entraron en vigencia las doctrinas aplicables a los Modelos de Inteligencia Artificial de Propósito General (GPAI). Exigen a los desarrolladores masivos implementar mitigaciones frente a riesgos algorítmicos sistémicos, transparencia algorítmica y la divulgación de las plantillas del corpus empleado, para sortear polémicas por infracciones de la propiedad intelectual mediante un Código de Prácticas. El ciclo concluye hacia agosto de 2026 y 2027, instanciando obligaciones monumentales para Sistemas de Alto Riesgo incrustados en infraestructuras (educación, contratación civil), exigiendo bitácoras inmutables de trazabilidad, validación cruzada y supervisión humana permanente para el etiquetado inequívoco de interfaces generativas e ilusiones multimedia sintéticas (deepfakes).

Esta cruzada preventiva ha provocado un debate estructural, llevando a multinacionales norteamericanas a frenar momentáneamente lanzamientos tecnológicos en la zona del euro, bajo temor a la represión burocrática, lo que motivó a la Comisión Europea a instaurar paquetes recientes de simplificación digital y de exenciones para PYMES.

Estados Unidos y la Desregulación para la Supremacía Tecnológica

En aguda contraposición a la UE, los Estados Unidos han abandonado doctrinas restrictivas a favor de un imperialismo tecnológico desregulado. En medio de un laberinto legislativo en el que las legislaturas estatales intentaban penalizar el sesgo corporativo, el gobierno federal estadounidense emitió en diciembre de 2025 una vigorosa Orden Ejecutiva presidencial (Aparicio, 2025).

Anulando mandatos precautorios previos, la Orden Ejecutiva impone la movilización del Departamento de Justicia para bloquear cautelarmente cualquier legislación fragmentada a nivel estatal que pretenda someter la innovación a

restricciones ideológicas o burocráticas disonantes con la priorización comercial federal. El fundamento es estratégico: liberar al sector privado del peso normativo para garantizar a las corporaciones estadounidenses la invulnerabilidad frente a competidores foráneos y cimentar la supremacía cibernética ininterrumpida.

China y la Férrea Disciplina Estatal Generativa

La República Popular China combina agresividad en la innovación con un control gubernamental totalitario sobre el ecosistema de la información. Fortaleciendo iterativamente sus ordenanzas desde mediados de 2023, la normativa vigente entre 2025 y 2026 somete la inteligencia artificial a la curaduría del Estado. Todos los Modelos Fundamentales y de Algoritmos de Recomendación están obligados a superar evaluaciones exhaustivas previas de integridad de seguridad socialista e inscribir sus repositorios ante la Administración del Ciberespacio de China (CAC).

Añadiendo presión técnica, los servicios que aglomeren a más de 1 millón de usuarios deben incorporar ineludiblemente marcas de agua, firmas visuales y acústicas y etiquetas criptográficas permanentes que prevengan la suplantación digital sintética ante el escrutinio de la población. En síntesis, las enmiendas a la matriz de la Ley de Ciberseguridad, programadas para su aplicación a partir del primero de enero de 2026, incorporan obligatoriamente los sistemas de redes neuronales como apéndices directos del mecanismo defensivo cibernético del Estado chino.

El Surgimiento de la Soberanía Algorítmica Global

Estados de considerable músculo económico que se rehúsan a ceder el monopolio geopolítico de la redacción lingüística corporativa occidental, como los Emiratos Árabes Unidos o Corea del Sur, están financiando centros de procesamiento propios, acumulando silos de silicio hardware (GPUs) independientes y

preentrenando LLMs encapsulados geográficamente, blindando herméticamente el filtrado de sus infraestructuras estratégicas, herencias culturales y arquitecturas militares, evitando subordinarse tanto a los controles estrictos de Europa como a los gigantes desregulados norteamericanos.

El vertiginoso despliegue cognitivo de los grandes modelos matemáticos ha transformado a la Inteligencia Artificial de ser una simple curiosidad técnica en erigirse como el factor reorganizador más significativo en la estructura económica global moderna; un proyecto colosal en el que la sola construcción de centros de datos subyugados eclipsa iniciativas históricas predecesoras (McKinsey & Company, 2024).

No obstante, las proyecciones para el cierre de la presente década oscilan entre dos polos radicalmente asimétricos. El primero subraya un riesgo de fragilidad teórica existencial: un estallido de burbuja motivado por el agotamiento asintótico de corpus lingüísticos vírgenes explotables y de barreras insalvables de la termodinámica operativa. Bajo este escenario sombrío, delineado por teóricos y académicos, los modelos generativos puros alcanzarían una meseta estancada, desprovistos de capacidad analítica a largo plazo en ausencia de una comprensión empírica del mundo tangible.

Por el lado opuesto, los líderes fundacionales de los centros de investigación más hegemónicos (Anthropic, OpenAI, xAI, Microsoft AI) auguran el punto crítico definitorio con una perturbadora cercanía. Visualizan la consecución ineludible de la Inteligencia General Artificial (AGI) —definida algorítmicamente como la capacidad de exhibir aptitud analítica e intervención técnica de nivel sobrehumano, automatizando tareas complejas multidimensionales como el diseño íntegro de software o flujos corporativos— concentrada dentro de una ventana de choque tan estrecha que se extendería apenas desde el año 2026 hasta finales del 2027.

La evidencia expuesta a lo largo del presente informe técnico concilia ambas perspectivas al concluir que la transición evolutiva decisiva hacia la AGI o Superinteligencia no radicará meramente en escalar infinitamente la dimensión de hiperparámetros de un transformador textográfico mediante pura fuerza bruta. Al contrario, dependerá umbilicalmente de una síntesis tecnológica convergente:

1. Dimensión Predictiva y Espacial: El desplazamiento irreversible de métodos predictivos estocásticos obsoletos (siguiente token) hacia las arquitecturas abstractas Modeladoras del Mundo Sensorial (V-JEPA, VL-JEPA), que decodifican intuitivamente las leyes gravitatorias y conceptuales, brindando agilidad logística en un plano físico a la robótica.
2. Epistemología Causal: La erradicación total de sesgos estadísticos basados en mera correlación en favor de motores analíticos imbuidos de Inferencia Causal, facultando al modelo artificial para aplicar razonamiento contrafactual abstracto, depurar covariables espurias y emitir resoluciones en ciencias exactas.
3. Hibridación Neuro-Simbólica Meta-Cognitiva: La soldadura matemática ineludible entre redes subsimbólicas fluidas (Sistema 1) y andamiajes de conocimiento ontológico explícito y de deducción lógica cristalizada (Sistema 2). El elemento vinculante faltante e indispensable consistirá en instaurar procesos dinámicos de auditabilidad introspectiva y de autocorrección del pensamiento.
4. Disrupción Física Biológica: Finalmente, quebrar el estancamiento y la entropía de los colapsos macroeléctricos mundiales mediante matrices matemáticas de extrema compresión (cuantización escalar binaria a 1 bit) y la consolidación definitiva del hardware neuromórfico exento de la herencia de von Neumann, posibilitando inteligencia ubicua a un costo en kilovatios homologado al de un cuerpo orgánico.

En retrospectiva crítica, el éxito sostenible en la próxima frontera técnica y

cognitiva no estará mediado únicamente por triunfos matemáticos puros. Quedará indisolublemente ligado a cuán audazmente el sector tecnológico logre un codiseño transparente e interpretable, resolviendo tensiones geopolíticas transcontinentales de soberanía algorítmica y, en última instancia, alineando la inmensurable omnisciencia del agente con la equidad innegociable de la sociedad humana en el siglo XXI.

Capítulo III
La Transformación Socioeconómica y los Imperativos Éticos de la Autonomía

La evolución contemporánea de la robótica se caracteriza por una transición fundamental desde sistemas rígidamente programados hacia agentes autónomos capaces de aprender de su entorno. Este cambio de paradigma se basa principalmente en el aprendizaje profundo (Deep Learning), una rama del aprendizaje automático que utiliza redes neuronales multicapa para modelar abstracciones de alto nivel a partir de datos. La integración de estas capacidades cognitivas en plataformas físicas ha dado lugar a una nueva generación de sistemas robóticos que no solo ejecutan tareas, sino que también perciben, razonan y actúan en entornos dinámicos y no estructurados.

Históricamente, el control robótico se basaba en modelos cinemáticos y dinámicos explícitos; sin embargo, la complejidad de las interacciones en el mundo real, como el contacto con objetos deformables o la navegación en condiciones de visibilidad degradada, ha superado las capacidades de la ingeniería tradicional. El aprendizaje profundo aborda estas limitaciones mediante arquitecturas que permiten procesar flujos de datos masivos y multimodales, lo que facilita tareas como la estimación de la profundidad, la navegación de extremo a extremo y la manipulación precisa de objetos.

3.1 Percepción Avanzada en Sistemas Robóticos

La percepción es el eslabón crítico inicial de la cadena de la autonomía. Los avances en los modelos de base visual (Visual Foundation Models) están transformando la localización y el mapeo robótico. En lugar de depender de descriptores de puntos específicos del dominio, los nuevos métodos aprovechan características aprendidas que son agnósticas a la estructura de la nube de puntos, lo que permite manejar de forma efectiva tanto escaneos LiDAR dispersos como mapas 3D densos, sin necesidad de reentrenamiento específico para cada entorno (Rodríguez et al., 2025).

Una innovación crítica en la navegación aérea es la reconstrucción de campos de radiancia neuronal (NeRFs) asistida por eventos. Para drones que vuelan a altas velocidades, el desenfoque por movimiento y el ruido en las estimaciones de pose representan serios obstáculos. La integración de flujos de eventos asíncronos con fotogramas convencionales permite recuperar campos de radiancia nítidos y trayectorias precisas sin una supervisión de verdad absoluta. Este marco de optimización conjunta refina la odometría visual-inercial basada en eventos y en modalidades de fotogramas, habilitando inspecciones rápidas de infraestructura y misiones de búsqueda y rescate en escenarios complejos.

La navegación robótica también se beneficia de la simulación diferenciable. Los algoritmos de gradiente de política de primer orden (FoPG) aprovechan la física de simulación local para acelerar la búsqueda de políticas, mejorando significativamente la eficiencia de muestreo en comparación con el aprendizaje por refuerzo libre de modelo estándar. Frameworks como RASH-BPTT (Real-world Anchored Short-horizon Backpropagation Through Time) permiten actualizaciones de políticas en vuelo robustas y eficientes, lo que permite a los cuadricópteros ejecutar maniobras ágiles cerca de los límites de saturación de sus actuadores. Estos hallazgos

subrayan que la adaptación en el mundo real no solo sirve para compensar errores de modelado, sino también como mecanismo práctico para la mejora sostenida del rendimiento en regímenes de vuelo agresivos (ver Tabla 7).

Tabla 7: Percepción avanzada en sistemas robóticos

Tecnología de percepción	**Mecanismo de Aprendizaje**	**Aplicación Principal**	**Ventaja Operativa**
NeRFs asistidos por eventos	Fusión asíncrona evento-imagen	Drones de alta velocidad	Eliminación del desenfoque por movimiento
Descriptores LiDAR aprendidos	Modelos de Base Visual	Mapeo y localización	Agnosticismo estructural y generalización
Percepción Activa	Regresión de navegación	Detección de objetos	Mejora de confianza mediante cambio de pose
SELM-SLAM3	SuperPoint + LightGlue	Navegación asistida	Robustez en escenas de baja textura
MANIP	Modularidad GOFE + Aprendido	Manipulación textil	Optimización de incertidumbre vs. recompensa

La arquitectura MANIP ejemplifica la tendencia actual de combinar métodos aprendidos con algoritmos procedimentales clásicos, conocidos como Good Old Fashioned Engineering (GOFE). Este enfoque modular permite integrar subpolíticas aprendidas para la percepción táctil o visual con primitivas algorítmicas establecidas, como la cinemática inversa, los filtros de Kalman y el control PID (Baleriola et al., 2025). Al emplear una metapolicía que determina qué módulo activar según un vector de confianza, los sistemas robóticos pueden alternar entre minimizar la incertidumbre

mediante la percepción interactiva y maximizar la recompensa durante la ejecución de la tarea. Esta integración es vital porque permite que el sistema se beneficie de la robustez de los algoritmos clásicos y aproveche la capacidad de adaptación de las redes neuronales ante datos sensoriales complejos.

Modelos Vision-Language-Action (VLA) e Inteligencia General Emboscada

El surgimiento de los modelos Vision-Language-Action (VLA) representa un hito en la creación de agentes robóticos generalistas. Estos modelos unifican la percepción visual, la comprensión del lenguaje natural y el control de dispositivos en un único marco de aprendizaje. A diferencia de los sistemas tradicionales fragmentados, donde la visión, el lenguaje y el control operan como silos independientes, los modelos VLA procesan simultáneamente las características visuales, el entendimiento semántico y las restricciones físicas del robot (Ge et al., 2025). La arquitectura de estos modelos suele basarse en transformadores (transformers), que emplean mecanismos de autoatención para procesar tokens de distintas modalidades de forma unificada.

El entrenamiento de estos modelos depende de conjuntos de datos masivos como Open X-Embodiment, que contiene cerca de un millón de trayectorias de robots de diversos tipos. Modelos como OpenVLA, con 7 mil millones de parámetros, han establecido nuevos estándares del arte al ser preentrenados con datos que abarcan múltiples configuraciones robóticas (generalización multiembodiment). Una de las decisiones arquitectónicas más críticas en el diseño de VLAs es cómo representar las acciones robóticas (véase la Tabla 8). Tradicionalmente, las acciones continuas se discretizan en tokens, lo que permite al modelo tratarlas como palabras de una oración. No obstante, enfoques emergentes como VLA-0 sugieren que representar acciones numéricas directamente como texto (strings) puede ser sorprendentemente potente, superando a modelos entrenados con datos a mayor escala sin necesidad de

cambios en la arquitectura base del modelo de lenguaje visual (VLM).

Tabla 8: Modelos Vision-Language-Action (VLA) e Inteligencia General Emboscada

Modelo VLA	Parámetros	Arquitectura Base	Innovación Clave
RT-1	N/A	Transformer Decoder	Mapeo directo multimodal a gran escala
RT-2	N/A	VLM (PaLM-E/ViT)	Coajuste fino con datos de internet
OpenVLA	7B	Llama 2 + SigLIP	Adaptabilidad mediante ajuste fino LoRA
π_0	N/A	Difusión / Latente	Control continuo y alta destreza
VLA-0	Variable	VLM Nativo	Representación de acciones como texto puro

A pesar de su promesa, los modelos VLA presentan limitaciones significativas en el razonamiento espacial y temporal en tareas de múltiples pasos. La integración de estos modelos en hardware real demanda una selección cuidadosa de la capacidad computacional, ya que el control en tiempo real a menudo requiere frecuencias de operación superiores a 200 Hz, lo que ha llevado al desarrollo de arquitecturas de sistema dual en las que un modelo lento procesa la semántica global y otro, rápido, ejecuta los comandos motores reactivos.

Aplicaciones de Vanguardia: Cirugía Autónoma y la Revolución de la Logística

La robótica, potenciada por el aprendizaje profundo, está penetrando en sectores en los que la precisión y la fiabilidad son imperativos absolutos. La medicina y la logística representan los dos polos de esta expansión tecnológica, impulsados por la necesidad de eficiencia y la escasez de mano de obra especializada.

El mercado de la robótica quirúrgica autónoma está experimentando una rápida expansión, con una proyección de crecimiento de $2.23 mil millones en 2024 a $2.61 mil millones en 2025. Esta evolución se sustenta en la creciente demanda de procedimientos mínimamente invasivos que prometen menores tiempos de recuperación, menor pérdida de sangre y una reducción general de las complicaciones postoperatorias. En especialidades como la urología y la ginecología, el uso de sistemas robóticos para procedimientos como la prostatectomía radical y la histerectomía ya representa entre el 25% y el 45% de todas las intervenciones en regiones avanzadas como los Estados Unidos.

La integración de la IA en estos sistemas permite pasar de la asistencia teleoperada a niveles incipientes de autonomía funcional. Algoritmos de visión computacional procesan tomografías preoperatorias para generar reconstrucciones 3D de la vasculatura del paciente, que se superponen en tiempo real sobre la vista del cirujano mediante realidad aumentada (AR) (Arguedas, 2024). Innovaciones como el sistema SASHA de Artedrone están diseñadas para realizar trombectomías mecánicas autónomas en pacientes con accidentes cerebrovasculares, navegando por la compleja red de vasos sanguíneos cerebrales para extraer coágulos con una precisión inalcanzable con el control manual. Además, la robótica blanda (soft robotics) permite desarrollar dispositivos implantables que cambian de forma para administrar fármacos de manera constante y evitar la fibrosis tisular.

A pesar de los beneficios, la transición hacia la autonomía plena en cirugía plantea retos de seguridad y de confianza. Actualmente, los sistemas se consideran no autónomos en el sentido legal, operando bajo la supervisión directa de un humano. El futuro apunta a sistemas capaces de identificar hitos anatómicos, evaluar la respuesta del tejido y guiar automáticamente las herramientas durante fases críticas, reduciendo así la dependencia del control manual y mitigando el impacto de la fatiga del cirujano, principal causa de errores quirúrgicos.

El giro Humano-Céntrico de la Industria 5.0

En los sectores logístico y de manufactura, se está produciendo un cambio filosófico desde la automatización orientada únicamente a la eficiencia (Industria 4.0) hacia un enfoque centrado en el ser humano, la resiliencia y la sostenibilidad (Industria 5.0) (ver Tabla 9). En este nuevo paradigma, los robots no se ven como sustitutos de la mano de obra, sino como colaboradores o amplificadores de capacidad que asumen tareas peligrosas, monótonas o físicamente exigentes (Da Costa et al., 2025).

La clave de esta colaboración reside en la capacidad de los robots para predecir la intención humana y adaptar su comportamiento en tiempo real. Esto se logra mediante el uso de gemelos digitales humanos (human digital twins), que modelan aspectos físicos, cognitivos y ergonómicos de los operarios para optimizar la interacción en el espacio de trabajo compartido. El uso de interfaces de lenguaje natural y de prompt engineering industrial permite que operarios sin conocimientos técnicos profundos de programación puedan instruir y coordinar sistemas robóticos complejos de manera intuitiva.

Tabla 9: Factor de diseño de la industria 4.0 y 5.0

Factor de Diseño	Industria 4.0 (Automatización)	Industria 5.0 (Colaboración)
Objetivo Primario	Eficiencia y rendimiento bruto	Bienestar humano y resiliencia
Rol del Humano	Supervisor del sistema	Socio creativo y tomador de decisiones
Entorno de trabajo	Robots en celdas cerradas	Espacios abiertos y compartidos
Tecnología Clave	Conectividad digital e IoT	IA predictiva y robots colaborativos
Medida de Éxito	Reducción de variabilidad	Flexibilidad y satisfacción del trabajador

La evidencia sugiere que los empleados que trabajan con robots colaborativos (cobots) reportan niveles más altos de autonomía y satisfacción laboral que quienes interactúan con sistemas industriales tradicionales, ya que se sienten valorados como orquestadores de la tecnología en lugar de ser meros operarios subordinados. Este cambio tiene implicaciones profundas para la resiliencia de las cadenas de suministro globales, que demostraron ser frágiles ante disrupciones repentinas; un sistema híbrido humano-robot es inherentemente más capaz de adaptarse a condiciones de mercado inciertas.

Impacto Socioeconómico y el Futuro del Trabajo

La proliferación de la robótica y el aprendizaje profundo suscita un debate polarizado sobre el desplazamiento laboral y la creación de nuevas oportunidades económicas. Las proyecciones para el horizonte 2024-2030 indican una reconfiguración masiva del mercado de trabajo global.

Los informes del Foro Económico Mundial sugieren que, si bien 85 millones de empleos podrían ser desplazados por la automatización para 2030, se espera la creación de 97 millones de nuevos puestos, lo que resultaría en un beneficio neto de 12 millones de empleos a nivel mundial. Sin embargo, estas cifras agregadas ocultan una realidad heterogénea. Los roles más expuestos a la automatización son aquellos basados en tareas repetitivas de baja calificación, como los trabajadores de las líneas de ensamblaje, los operadores de máquinas y el personal de embalaje. En contraste, la demanda de trabajadores del conocimiento, especialistas en datos y diseñadores de sistemas de IA, está en aumento, lo que crea una estructura ocupacional piramidal en la que los niveles más bajos de capital humano enfrentan un riesgo existencial de exclusión.

La velocidad de este desplazamiento es un factor crítico. A diferencia de las olas tecnológicas anteriores que se desarrollaron a lo largo de décadas, la combinación de robots humanoides y modelos de IA avanzados podría transformar sectores enteros en un plazo de 5 a 10 años, dejando poco tiempo para que los sistemas educativos y los marcos de política laboral se adapten. Además, existe el riesgo de que los beneficios de productividad derivados de la automatización se concentren en los propietarios del capital, lo que podría exacerbar la desigualdad de ingresos si no se implementan medidas de redistribución activa ni se imponen impuestos a la automatización (Pérez, 2024).

El Desafío de la Polarización y la Movilidad Ocupacional

Un efecto secundario preocupante de la automatización es la reducción de la movilidad profesional entre los trabajadores de baja calificación. Las investigaciones indican que cuando se introduce un robot adicional por cada 1.000 trabajadores en un mercado local, el valor de la carrera de por vida de los trabajadores afectados disminuye significativamente, ya que encuentran dificultades para acceder a empleos

con salarios similares o superiores. Esta falta de opciones reduce su poder de negociación salarial y puede desencadenar un efecto dominó en las comunidades locales, donde la disminución de los ingresos se traduce en una menor inversión en vivienda y educación.

En regiones con alta dependencia industrial, como el Cinturón del Óxido en EE. UU., este declive económico se ha correlacionado con un cambio en las preferencias políticas hacia agendas populistas, lo que sugiere que el impacto de la robótica trasciende lo económico para influir en la estabilidad democrática. Por otro lado, en economías con poblaciones envejecidas, la robótica no se percibe como una amenaza de desempleo, sino como una solución vital ante la escasez de trabajadores en edad productiva para sostener los servicios básicos y la producción industrial.

La respuesta institucional a estos cambios debe centrarse en programas de reentrenamiento proactivos (reskilling). Las organizaciones que invierten en la formación continua de sus empleados logran retener el conocimiento institucional mientras se adaptan al cambio tecnológico. El auge de los bootcamps de ciencia de datos y de programas de credencialización rápida demuestra que es posible adquirir nuevas habilidades en meses en lugar de años, con tasas de empleo posformación del 70 al 80%.

Imperativos Éticos

La falta de transparencia en la toma de decisiones robóticas resulta particularmente peligrosa en aplicaciones de alto impacto. Los algoritmos pueden heredar y amplificar prejuicios humanos presentes en los datos de entrenamiento. Por ejemplo, se han documentado sistemas de IA que bajo-diagnostican a ciertos grupos demográficos o que utilizan el origen étnico para predecir reincidencias delictivas de manera discriminatoria (Araya, 2021). En la robótica quirúrgica, la opacidad de los

algoritmos de diagnóstico o de asistencia puede llevar a errores imprevistos que ni el cirujano ni el fabricante pueden identificar con precisión por qué el sistema tomó una acción dañina.

Para combatir estos riesgos, la comunidad técnica está impulsando el concepto de Ética por Diseño. El IEEE ha desarrollado el marco de Diseño Alineado Éticamente (Ethically Aligned Design), que establece principios fundamentales para el desarrollo de sistemas autónomos e inteligentes que prioricen los derechos humanos y el bienestar (eudaimonía).

Principios Generales del Diseño Ético (IEEE):

1. **Derechos Humanos:** Los sistemas no deben infringir las libertades fundamentales reconocidas internacionalmente.
2. **Bienestar:** Los creadores deben adoptar métricas de calidad de vida como criterio principal de éxito, por encima del crecimiento comercial.
3. **Agencia de Datos:** Empoderar a los individuos para ejercer el control sobre su identidad y sus datos personales.
4. **Transparencia:** Las decisiones deben ser siempre descubribles y explicables.
5. **Responsabilidad:** Proporcionar una justificación clara y rastreable para todas las acciones del sistema.

Dilemas de Responsabilidad Civil y Legal

La autonomía robótica desafía los marcos legales tradicionales de responsabilidad. En un accidente de vehículo autónomo, la culpa puede distribuirse de manera compleja entre el fabricante del vehículo, el desarrollador del software, el operador de la flota o incluso el propietario del vehículo si descuidó el mantenimiento de los sensores.

En la medicina, la jurisprudencia se debate sobre si las fallas de un robot

quirúrgico deben tratarse bajo el régimen de la mala praxis médica o bajo el de la responsabilidad por producto. Si el cirujano operó el dispositivo incorrectamente, la responsabilidad es personal; sin embargo, si el fallo se debió a un bug de software o a un error en el reconocimiento de objetos, el fabricante podría ser hallado responsable por negligencia en el diseño o la fabricación (Shentu, 2024). La doctrina del Intermediario Aprendido (Learned Intermediary) complica aún más este panorama, al absolver, en ocasiones, al fabricante si este proporcionó advertencias adecuadas al médico, trasladando a este último la obligación de informar al paciente sobre los riesgos específicos de la tecnología.

La aparición de sistemas con capacidad de aprendizaje continuo introduce un riesgo adicional: los robots pueden evolucionar más allá de la previsión original de sus fabricantes, por lo que resulta imperativo que las agencias reguladoras, como la FDA o la NHTSA, establezcan esquemas de responsabilidad proactivos antes de que la tecnología se adopte de forma masiva.

Gobernanza y Regulación Internacional: El Modelo del EU AI Act

La Unión Europea ha asumido el liderazgo global en la regulación de la IA con la aprobación del EU AI Act en 2024. Este marco legal adopta un enfoque basado en el riesgo para clasificar los sistemas de IA y determinar sus obligaciones de cumplimiento.

Clasificación por Niveles de Riesgo

El reglamento distingue cuatro categorías principales:

- **Riesgo Inaceptable:** Incluye sistemas que manipulan el comportamiento humano, la puntuación social por parte de los gobiernos y la vigilancia masiva en tiempo real. Estos sistemas están prohibidos en la UE.
- **Alto Riesgo:** Sistemas utilizados en infraestructuras críticas, dispositivos

médicos (incluidos robots quirúrgicos), gestión del empleo y de la educación. Estos deben someterse a evaluaciones rigurosas de conformidad, implementar sistemas de gestión de riesgos y garantizar la supervisión humana.

- **Riesgo Limitado:** Sistemas como chatbots o deepfakes que deben cumplir con obligaciones de transparencia, informando a los usuarios que interactúan con una IA.
- **Riesgo Mínimo:** Incluye aplicaciones como filtros de spam o videojuegos con IA, que no están sujetos a regulaciones adicionales bajo esta ley.

Los proveedores de robots de alto riesgo deben establecer un proceso de gestión de riesgos iterativo y continuo a lo largo de todo el ciclo de vida del producto. Esto incluye el análisis de riesgos previsibles para la salud, la seguridad y los derechos fundamentales, así como la realización de pruebas en condiciones reales antes de la comercialización (OECD/CAFD, 2022). La ley también exige la trazabilidad de las decisiones, obligando a los sistemas a registrar automáticamente los eventos para permitir la investigación de fallos. Las multas por incumplimiento son severas y pueden alcanzar hasta el 7% de la facturación global de una empresa, lo que subraya la seriedad con que se aborda la gobernanza de la IA.

3.2 El Ecosistema de la Robótica e IA en Perú: Avances y Estrategia Nacional

Perú ha emergido como un actor proactivo en la región latinoamericana en el desarrollo de políticas de IA y robótica. El país ha pasado de una fase de adopción informal a la creación de un marco normativo y estratégico robusto.

Marco Legal: La Ley 31814 y su Reglamento

En julio de 2023, se promulgó la Ley N° 31814, que promueve el uso de la IA para el desarrollo económico y social del país, bajo los principios de ética,

transparencia y seguridad basada en riesgos. En septiembre de 2025, el reglamento de esta ley fue oficializado mediante el Decreto Supremo N° 115-2025-PCM, lo que estableció hitos claros para el desarrollo tecnológico peruano.

El reglamento clasifica los usos de la IA de manera similar al modelo europeo, prohibiendo la generación de capacidad letal autónoma en el ámbito civil, la manipulación engañosa de personas y el perfilamiento discriminatorio basado en datos sensibles. Asimismo, exige la supervisión humana obligatoria en sistemas de alto riesgo, definidos como aquellos que afectan la vida, la dignidad y la seguridad física de los ciudadanos, como el software de evaluación crediticia o los diagnósticos médicos asistidos por ordenador. Este enfoque busca equilibrar el impulso a la innovación con la protección de los derechos fundamentales, un equilibrio crítico en un país con brechas sociales significativas.

Estrategia Nacional de IA (ENIA) y Talento Digital

La Estrategia Nacional de Inteligencia Artificial (ENIA), liderada por la Secretaría de Gobierno y Transformación Digital (SGTD-PCM), proyecta metas hasta el año 2030. Entre sus pilares se encuentran la promoción del talento digital y la creación de un Centro Nacional de Innovación e Inteligencia Artificial. La estrategia reconoce que la IA adaptada a las necesidades de los estudiantes podría individualizar el aprendizaje desde la etapa preescolar hasta la universidad, optimizando la educación nacional.

En el sector productivo, se están explorando aplicaciones de IA y robótica en la agricultura para reducir el consumo de agua y energía y en la minería para optimizar procesos y mejorar la seguridad laboral. El mercado peruano de robótica ha mostrado un dinamismo notable, expandiéndose un 20% en 2025 debido a la demanda de soluciones de seguridad y automatización por parte de universidades y pymes.

Investigación y Académica en el Perú

Las universidades peruanas desempeñan un papel central en la generación de conocimiento y en la ética de la robótica.

- **Universidad Nacional Mayor de San Marcos (UNMSM):** Ha consolidado la robótica como una disciplina estratégica, al inaugurar una carrera profesional en IA y al organizar eventos internacionales, como la Escuela Latinoamericana de Inteligencia Artificial (LASAI). Además, fomenta la innovación estudiantil a través de torneos como La Decanatron y de su incubadora de empresas 1551.
- **Pontificia Universidad Católica del Perú (PUCP):** A través del Grupo de Investigación en Robótica Aplicada y Biomecánica (GIRAB), desarrolla tecnologías para la rehabilitación física y el análisis del movimiento humano. El grupo es interdisciplinario e integra a expertos en ingeniería, medicina y psicología para diseñar prótesis mioeléctricas e interfaces cerebro-computador.
- **Iniciativas Interuniversitarias:** Estudiantes de la UNMSM, PUCP y UNAC han colaborado en el diseño de robots humanoides y sistemas móviles para aplicaciones de rescate y seguridad, posicionando a Lima como un epicentro regional de innovación robótica (ver Tabla 10).

Tabla 10: Acción e impacto generado en investigación académica desde las universidades peruanas

Institución / Iniciativa	Foco de Investigación / Acción	Impacto Generado
GIRAB-PUCP	Biomecánica y rehabilitación	Prótesis avanzadas y análisis de marcha
LASAI (San Marcos)	Capacitación en IA y ML	Formación de investigadores regionales
Incubadora 1551	Emprendimiento tecnológico	Formalización de startups de base científica

Ley 31814	Marco Regulatorio	Uso ético y responsable de la IA en Perú
Glexco Robotics	Mercado Industrial	Expansión del 20% en soluciones robóticas

A pesar de estos avances, el ecosistema peruano enfrenta desafíos en materia de infraestructura digital en zonas rurales y en la necesidad de una mayor inversión en investigación y desarrollo (I+D) que permita pasar de la adaptación de tecnologías extranjeras a la creación de propiedad intelectual propia.

La convergencia entre el aprendizaje profundo y la robótica marca el inicio de una era de autonomía inteligente que promete beneficios transformadores, pero exige una vigilancia ética constante. El análisis de las tendencias actuales permite inferir varios escenarios para el final de la década.

Primero, la transición hacia la Industria 5.0 consolidará la visión de los robots como socios y no como competidores del ser humano. El éxito de las empresas no se medirá solo por su tasa de automatización, sino por su capacidad para integrar el juicio creativo humano con la precisión de la máquina de manera resiliente. Segundo, la madurez de los modelos Vision-Language-Action (VLA) permitirá que los robots salgan de los entornos controlados de las fábricas para operar de manera efectiva en hogares, hospitales y espacios públicos, democratizando el acceso a la asistencia tecnológica.

Tercero, la gobernanza de la IA se volverá cada vez más fragmentada geográficamente a menos que se alcancen consensos internacionales sobre principios éticos universales. El modelo del EU AI Act servirá de referencia, pero países como Perú demuestran que las regulaciones locales deben adaptarse a contextos sociales específicos, como la informalidad económica y la brecha digital. Cuarto, la educación

y el reentrenamiento laboral serán los pilares de la paz social; la inversión en capital humano es la única salvaguardia efectiva contra la polarización económica derivada de la automatización (Llacsa et al., 2026).

En conclusión, la robótica potenciada por el aprendizaje profundo no es un fenómeno puramente técnico, sino una reconfiguración de la relación entre la humanidad y sus herramientas. El imperativo para los próximos años es garantizar que esta autonomía creciente se traduzca en un aumento genuino de la calidad de vida, protegiendo la dignidad humana frente a la opacidad de los algoritmos y asegurando que los beneficios de la inteligencia artificial se distribuyan de manera equitativa en toda la sociedad.

Capítulo IV
Gobernanza Global y Soberanía de la Inteligencia Artificial

La arquitectura del poder internacional en marzo de 2026 se encuentra en una profunda metamorfosis, impulsada por la integración sistémica de la inteligencia artificial (IA) en las estructuras de defensa, de economía y de cohesión social. Lo que en años anteriores se percibía como una competencia tecnológica sectorial ha cristalizado ahora en una lucha existencial por la autonomía estratégica y el control de los marcos éticos que regirán el comportamiento de los sistemas autónomos en la próxima década. La dialéctica entre la gobernanza global y la soberanía tecnológica no es simplemente un debate legal, sino la manifestación de una nueva geografía del poder donde el control de los algoritmos define la frontera entre la innovación y la dominación.

4.1 El Nuevo Paradigma de la Gobernanza Global: De los Principios a la Operatividad

La gobernanza de la IA ha trascendido las declaraciones de intención para convertirse en un terreno de confrontación y de cooperación regulada. Por ende, la noción de que un Estado o grupo de Estados puede controlar unilateralmente la tecnología ha sido descartada ante la evidencia de su naturaleza transfronteriza por esencia (Estrada, 2025). La infraestructura de la IA, desde la extracción de minerales críticos hasta el refinamiento de datos de entrenamiento, se distribuye globalmente, lo que exige un enfoque de red ágil, flexible y, sobre todo, inclusivo para evitar que el Sur Global quede permanentemente al margen del desarrollo tecnológico.

El liderazgo de las Naciones Unidas y el diálogo global

En el marco del informe Governing AI for Humanity, las Naciones Unidas han establecido que la gobernanza de la IA no debe desarrollarse en un vacío, sino que debe anclarse firmemente en el derecho internacional y en los derechos humanos. Este enfoque busca mitigar los riesgos asociados a las alucinaciones de los grandes modelos lingüísticos (LLM), la desinformación a escala masiva y las amenazas a la privacidad, al tiempo que garantiza que los beneficios de la IA contribuyan al logro de los Objetivos de Desarrollo Sostenible (ODS) (Naciones Unidas, 2024).

Así, el Diálogo Global sobre Gobernanza de la IA, respaldado por la ONU, y el Panel Científico Internacional Independiente sobre IA constituyen los pilares de este esfuerzo multilateral. Por primera vez, casi todos los Estados cuentan con un foro para debatir normas de transparencia y mecanismos de coordinación, con el fin de equilibrar la asimetría informativa entre las corporaciones que poseen laboratorios de vanguardia y el resto del mundo. No obstante, la efectividad de este marco es frágil, ya que las potencias suelen evitar compromisos vinculantes en áreas de alto riesgo como la vigilancia masiva o el despliegue de armas autónomas.

La Postura del Sur Global frente a la Dominación Algorítmica

Los líderes de las economías emergentes han intensificado su retórica contra lo que denominan dominación algorítmica. El presidente Luiz Inácio Lula da Silva ha propuesto una gobernanza global bajo el mando directo de la ONU, argumentando que cuando una minoría controla los algoritmos, el resultado no es progreso, sino una nueva forma de explotación. Esta visión es compartida por bloques regionales que buscan que la IA reconozca la diversidad de las trayectorias nacionales y fortalezca la soberanía en lugar de erosionarla (ver Tabla 11).

Tabla 11: Principio Rector de Gobernanza (ONU 2026)

Principio Rector de Gobernanza (ONU 2026)	Descripción y Objetivo
Inclusividad Global	La IA debe utilizarse en beneficio de todos, reduciendo la brecha entre desarrolladores y usuarios.
Marco de Derechos Humanos	Las decisiones en el ciclo de vida de la IA deben respetar la dignidad, la privacidad y la no discriminación.
Interoperabilidad Normativa	Convertir las iniciativas locales en un todo coherente basado en el derecho internacional.
Transparencia Científica	Acceso a información imparcial para equilibrar las ventajas de los laboratorios corporativos.

4.2 La Soberanía de la IA como Eje de la Resiliencia Estratégica

En la actualidad, la IA soberana se define como la capacidad de una nación para influir, desarrollar y desplegar tecnología de acuerdo con sus intereses nacionales y valores culturales, evitando dependencias insostenibles de actores externos. Esta autonomía ya no se mide únicamente en términos de desarrollo de software, sino también en la propiedad y el control de la pila tecnológica (tech stack), que incluye el poder de cómputo, la infraestructura en la nube y el acceso a semiconductores avanzados.

La Brecha Computacional y la Bipolaridad Tecnológica

El panorama de la IA está marcado por una profunda brecha entre los líderes (EE. UU. y China) y el resto del mundo. A marzo de 2025, Estados Unidos concentraba el 75% del rendimiento global de las supercomputadoras de IA, mientras que China poseía el 15%. Esta concentración ha relegado a potencias industriales tradicionales como Alemania, Japón y el Reino Unido a un papel secundario, obligándolas a

redefinir sus estrategias de supervivencia tecnológica.

Las potencias medias enfrentan el riesgo de quedar bloqueadas en ecosistemas tecnológicos que no controlan, lo que compromete su capacidad para gestionar los servicios públicos, la defensa y el crecimiento económico. La dependencia de modelos de frontera propietarios, como los desarrollados por Google, Microsoft o OpenAI, o de modelos de código abierto avanzados de empresas chinas como DeepSeek o Alibaba, genera una vulnerabilidad estructural que los Estados buscan mitigar mediante diversas estrategias de adaptación (Naciones Unidas, 2024).

Estrategias de las Potencias Medias para el 2026

Ante la imposibilidad de igualar la escala de inversión de las superpotencias, los países intermedios han adoptado enfoques pragmáticos descritos por analistas de riesgo geopolítico:

- **Especialización:** Países que eligen un nicho específico en la cadena de suministro global, como Arabia Saudita, con su modelo METABRAIN, enfocado exclusivamente en la industria energética.
- **Alineación:** La formación de asociaciones bilaterales profundas con una superpotencia para garantizar el suministro de infraestructura, como los pactos de EE. UU. con los Emiratos Árabes Unidos en 2025.
- **Soberanía Compartida:** Iniciativas para agrupar recursos y crear infraestructuras públicas digitales comunes, ejemplificadas por el Citizen Stack de la India o por los proyectos de computación de alto rendimiento en Europa.
- **Cobertura (Hedging):** Una combinación estratégica de capacidades nacionales y de asociaciones con múltiples proveedores extranjeros para evitar la coacción y garantizar la continuidad operativa.

Implementación del Reglamento de IA de la Unión Europea (EU AI Act)

El Reglamento de IA de la UE se ha convertido en el estándar de facto para la regulación global, influyendo en las legislaciones de múltiples países a través del efecto Bruselas. A marzo de 2026, el reglamento se encuentra en una fase crítica de aplicación gradual, con hitos significativos que afectan a empresas de todo el mundo que interactúan con el mercado único europeo.

Estatus y Obligaciones de Cumplimiento

Desde febrero de 2025, las prohibiciones sobre sistemas de IA que presentan riesgos inaceptables, como el social scoring o la manipulación psicológica, son plenamente vinculantes. Para agosto de 2025, entraron en vigor las reglas para los modelos de IA de propósito general (GPAI) y los sistemas de gobernanza nacional. Sin embargo, la plena aplicabilidad de la mayoría de las disposiciones restantes, especialmente las relativas a los sistemas de alto riesgo, queda fijada para el 2 de agosto de 2026 (ver Tabla 12).

Tabla 12: Categoría de riesgo y transparencia de modelos de IA

Categoría de IA (Reglamento UE)	Ejemplos de aplicación	Requisitos de Cumplimiento en 2026
Riesgo Inaceptable	Puntuación social, vigilancia biométrica masiva.	Prohibición total; la infracción conlleva multas máximas.
Alto Riesgo	Evaluación crediticia, selección de personal, infraestructuras críticas.	Registro en la base de datos de la UE, documentación técnica y supervisión humana.
Transparencia	Chatbots, contenido generado sintéticamente (deepfakes).	Etiquetado claro de contenido artificial y notificación al usuario.

Modelos GPAI	Modelos de lenguaje de propósito general (p. ej., Claude, GPT-4).	Cumplimiento de los derechos de autor y publicación de resúmenes de entrenamiento.

Desafíos en la Estandarización y Guías Técnicas

Un desafío notable en marzo de 2026 es el retraso en la publicación de las guías definitivas por parte de la Comisión Europea. El plazo de febrero de 2026 para proporcionar directrices sobre la implementación del Artículo 6 (determinación de sistemas de alto riesgo) fue incumplido, lo que ha generado incertidumbre en el sector empresarial. Los organismos de estandarización europeos (CEN y CENELEC) también han pospuesto la publicación de normas técnicas hasta finales de 2026, lo que dificulta la realización de evaluaciones de conformidad precisas por parte de las PYMES y startups.

4.3 El Caso de Perú: Marco Normativo y Estrategia 2026-2030

Perú ha emergido como un referente regional en la gobernanza de la IA, siendo uno de los primeros países de América Latina en establecer un marco legal detallado que busca equilibrar el fomento económico con la protección de los derechos fundamentales (Estrada, 2025).

La Ley N° 31814 y su Reglamento (2025-2026)

La Ley N° 31814, publicada en 2023, sentó las bases para el uso ético y responsable de la IA. El 9 de septiembre de 2025, se aprobó el Reglamento de dicha ley mediante el Decreto Supremo 115-2025-PCM, que entró en vigencia el 22 de enero de 2026. Este reglamento clasifica el uso de la IA en tres niveles de riesgo: uso indebido (prohibido), uso de alto riesgo (sujeto a controles estrictos) y uso aceptable.

Estrategia Nacional de IA y Soberanía Digital

La Propuesta de la Estrategia Nacional de Inteligencia Artificial 2026-2030 tiene como objetivo general fortalecer el ecosistema nacional mediante la formación de talento humano, el impulso de la infraestructura de datos y la creación de un Centro Nacional de Innovación e Inteligencia Artificial. Un componente vital de esta estrategia es la soberanía de datos, que asegura que los activos digitales estratégicos del Estado se utilicen de manera ética para impactar en el desarrollo sostenible, sin generar dependencias tecnológicas irreversibles (Smuha, 2025).

El informe RAM de la UNESCO (Metodología de Evaluación del Estado de Preparación), actualizado en enero de 2026, destaca que, si bien Perú lidera en gobernanza digital, aún enfrenta desafíos en la capacidad de cómputo y en la necesidad de evitar sesgos algorítmicos que puedan afectar a poblaciones indígenas o rurales, debido a la falta de datos representativos de su diversidad cultural.

El Conflicto Anthropic-Pentágono: Un Punto de Inflexión en la Ética Militar

En febrero y marzo de 2026, la relación entre el Estado y las corporaciones tecnológicas experimentó una crisis sin precedentes que ha redefinido los límites de la soberanía tecnológica y de la autonomía corporativa. El conflicto entre la empresa Anthropic y el Departamento de Defensa de EE. UU. (rebautizado como Departamento de Guerra) ilustra la tensión entre los marcos éticos privados y las exigencias de la seguridad nacional.

El 28 de febrero de 2026, la administración Trump prohibió el uso federal de la tecnología de Anthropic después de que la empresa se negara a eliminar restricciones éticas en su modelo Claude para dos aplicaciones específicas: las armas autónomas letales (LAWS) y la vigilancia masiva de ciudadanos estadounidenses. Anthropic argumentó que su compromiso con la seguridad de la IA y con los valores

democráticos impedía que su tecnología se utilizara en sistemas que eliminaran el control humano en decisiones de vida o muerte.

La respuesta del Pentágono fue contundente. El secretario de Guerra, Pete Hegseth, designó a Anthropic como un riesgo para la cadena de suministro de seguridad nacional, una clasificación históricamente reservada para entidades vinculadas a adversarios extranjeros. Además, se amenazó con prohibir cualquier relación comercial entre contratistas de defensa y Anthropic, lo que representaría una sentencia de muerte comercial para la empresa.

Tendencias de Ciberseguridad y Transparencia

La ciberseguridad en la era de la IA ha dejado de ser una función de soporte para convertirse en el núcleo de la resiliencia nacional. Hoy, el concepto de perímetro tradicional ha desaparecido y ha sido reemplazado por la identidad y el acceso como nuevas fronteras de protección (Linares, 2024).

MLSecOps y la Protección de Modelos

Las organizaciones más maduras han adoptado el enfoque MLSecOps (Machine Learning Security Operations). Ya no basta con proteger los servidores; es necesario proteger los datos de entrenamiento contra el envenenamiento (data poisoning), evitar el robo de pesos de los modelos (model stealing) y monitorizar activamente las inyecciones de prompts que podrían exfiltrar información confidencial.

En 2026, el 80% de las intrusiones avanzadas involucran la explotación de credenciales. La arquitectura Zero Trust se ha vuelto imperativa, con la implementación de autenticación multifactor robusta y segmentación basada en identidad para limitar el movimiento lateral de los atacantes en las redes corporativas

y gubernamentales.

IA Generativa 2.0 en la Defensa y el Ataque

La ciberseguridad se ha convertido en una carrera armamentista de algoritmos. Los atacantes utilizan modelos de IA para crear malware bajo demanda que se adapta en tiempo real al entorno de la víctima. Por el contrario, los Centros de Operaciones de Seguridad (SOC) inteligentes utilizan IA para correlacionar señales dispersas, identificar patrones anómalos y ejecutar respuestas de contención autónomas basadas en playbooks preentrenados (Okdem & Okdem, 2024).

Un avance significativo es la virtualización del análisis forense. En entornos de tecnología operacional (OT) críticos, como plantas de energía o sistemas de transporte, la IA permite crear imágenes forenses y probar parches en entornos aislados antes de su aplicación, garantizando que la respuesta a incidentes no interrumpa procesos industriales vitales.

4.4 Proyecciones Económicas y Sociales hacia el 2030

El impacto de la IA no se limita a la seguridad y a la geopolítica; está reconfigurando las bases mismas de la economía global y de la identidad cultural. Para el año 2030, las proyecciones indican una transformación estructural sin precedentes.

El Estímulo Económico y la Brecha de Productividad

Se estima que la IA contribuirá con aproximadamente 19,9 billones de dólares a la economía global hasta 2030, lo que representará el 3,5 % del PIB mundial para esa fecha. Este crecimiento se impulsará por ganancias masivas de productividad en los sectores adoptantes y por la expansión de la cadena de suministro de servicios de IA. Sin embargo, este beneficio no será uniforme. Los países con mercados laborales más

formales y un acceso digital más amplio, como algunos de América Latina, tienen mayores probabilidades de capitalizar estas ventajas.

El Fondo Monetario Internacional (FMI) advierte que la IA afectará al 40% del empleo mundial. Mientras que en las economías avanzadas la IA puede complementar trabajos altamente cualificados, en las naciones en desarrollo existe el riesgo de que profundice la desigualdad si no se establecen redes de seguridad social y programas de reentrenamiento (*reskilling*) para los trabajadores vulnerables.

Sostenibilidad y los Objetivos de Desarrollo Sostenible (ODS)

La relación entre la IA y la sostenibilidad es una espada de doble filo. Por un lado, la IA tiene el potencial de optimizar el consumo de energía y acelerar la transición hacia fuentes renovables (ODS 7). Por otro lado, el entrenamiento de modelos de IA a gran escala consume cantidades ingentes de energía y agua para la refrigeración de los centros de datos (Garikapati et al., 2025).

Esto implica no solo usar energía limpia para el cómputo, sino también diseñar algoritmos más eficientes desde su concepción. Empresas líderes han comenzado a rastrear la huella de carbono de sus cadenas de suministro globales mediante herramientas de IA generativa, asegurando la transparencia y el cumplimiento de las normativas de sostenibilidad cada vez más estrictas.

El análisis de la gobernanza global y de la soberanía de la IA en el horizonte revela una tensión irresoluble pero gestionable. La búsqueda de la soberanía total es una ilusión que desaparece, ya que ninguna nación puede ser completamente autosuficiente en el ecosistema global de semiconductores, datos y talento. Sin embargo, la dependencia total constituye un riesgo estratégico inaceptable. La respuesta que emerge es la autonomía selectiva o la interdependencia estratégica. Los países que logren prosperar en 2026 no serán necesariamente los que desarrollen los

modelos más grandes, sino aquellos que tengan la capacidad de:

1. **Gobernar sus propios datos:** Implementar marcos, como el MIGDIA, p. ej., la OEA, para asegurar que la información nacional se utilice para el bien público y no sea extraída sin beneficio local.
2. **Regular con flexibilidad:** Adoptando marcos como el de Perú o el de la UE que mitigan riesgos sin ahogar la innovación de las PYMES.
3. **Construir coaliciones:** Actuando colectivamente en foros como las Naciones Unidas o en bloques regionales para influir en las normas globales y evitar la bipolaridad absoluta entre EE. UU. y Rusia, EE. UU. y China.
4. **Mantener el control humano:** Garantizar que, en las aplicaciones de alto riesgo, especialmente en la justicia, la salud y la defensa, la decisión final siempre resida en una autoridad humana ética y responsable.

En última instancia, la soberanía de la IA no consiste en el aislamiento, sino en la capacidad de elegir, adaptar y dirigir la tecnología para el bienestar de la población y la preservación de los valores democráticos en un mundo irremediablemente conectado.

Capítulo V
El Giro Humano-Céntrico de la Industria 5.0: Tecnología, Sociedad y Resiliencia

La evolución de la producción industrial ha alcanzado un estadio en el que la eficiencia técnica y la automatización ya no bastan para sostener el progreso sistémico de la sociedad contemporánea. La transición hacia la Industria 5.0 representa una reorientación fundamental de los paradigmas productivos, desplazando el enfoque tecnocéntrico de la Industria 4.0 hacia una visión en la que el bienestar del trabajador, la sostenibilidad ambiental y la resiliencia organizacional constituyen los ejes rectores de la innovación. Según la Comisión Europea, este nuevo marco no busca reemplazar los avances de la digitalización, sino subordinar el desarrollo tecnológico a los límites biofísicos del planeta y a las necesidades intrínsecas del ser humano, y posicionar a la industria como un motor de prosperidad que trasciende el simple crecimiento del Producto Interno Bruto.

Este cambio de paradigma se sustenta en el reconocimiento de que la digitalización masiva, aunque ha optimizado la productividad, a menudo ha descuidado el factor humano, relegando al trabajador a roles de supervisión pasiva y generando una brecha de habilidades que amenaza la cohesión social y la estabilidad operativa. La Industria 5.0 emerge, por tanto, como una respuesta estratégica a las crisis globales —incluyendo la pandemia de COVID-19, las tensiones geopolíticas y la emergencia climática— que han evidenciado la fragilidad de las cadenas de suministro hipereficientes pero rígidas. En este contexto, la resiliencia se redefine no solo como la capacidad de sobrevivir a una crisis, sino también como la facultad de

antifragilidad, que permite que los sistemas industriales aprendan y se fortalezcan a través de la disrupción.

5.1 Evolución Conceptual: De la Automatización a la Colaboración Simbiótica

La trayectoria de las revoluciones industriales refleja una búsqueda constante de mayor control sobre el entorno productivo. Mientras que la Industria 1.0 introdujo la mecanización, la 2.0 la producción en masa y la 3.0 la electrónica, la Industria 4.0 se caracterizó por la integración de sistemas ciberfísicos, el Internet de las Cosas (IoT) y la inteligencia artificial para crear fábricas inteligentes autónomas. Sin embargo, la Industria 5.0 introduce un cambio cualitativo al reintegrar la creatividad humana y la artesanía en el proceso industrial, utilizando la tecnología para potenciar las capacidades cognitivas y físicas del operario en lugar de sustituirlas (Da Costa et al., 2025).

La diferencia fundamental radica en la intencionalidad estratégica. La Industria 4.0 se centró en la velocidad y la precisión mediante la automatización, lo que a menudo derivó en la desvinculación del trabajador. En contraste, la Industria 5.0 promueve la sinergia y el propósito, en la que los sistemas inteligentes gestionan las tareas peligrosas, monótonas o físicamente agotadoras (las denominadas tareas 3D: *dirty, dangerous, dull*), lo que permite que los humanos se centren en la innovación, el pensamiento crítico y la toma de decisiones complejas. Este enfoque se resume en el concepto del Triple Resultado (Triple Bottom Line), que equilibra las Personas, el Planeta y los Beneficios (*People, Planet, Profit*).

Tabla 13: Comparativa analítica de paradigmas industriales

Atributo Estratégico	Industria 4.0 (Tecnocéntrica)	Industria 5.0 (Humanocéntrica)
Núcleo de Valor	Eficiencia operativa y productividad.	Bienestar humano y valor social.
Rol de la tecnología	Automatización total y autonomía de las máquinas.	Aumento humano y colaboración (Cobots).
Sostenibilidad	Mitigación de riesgos y cumplimiento.	Enfoque neto positivo y circularidad.
Estructura del Sistema	Sistemas ciberfísicos integrados.	Ecosistemas socio-técnicos viables.
Flexibilidad	Personalización masiva basada en datos.	Personalización extrema y artesanía digital.
Resiliencia	Robustez basada en la estabilidad de los datos.	Adaptabilidad y aprendizaje ante la crisis.

La transición hacia la Industria 5.0 no implica el abandono de la infraestructura digital de la versión 4.0. Por el contrario, utiliza los sensores conectados, el análisis de

grandes datos y la infraestructura de la nube como condición sine qua non sobre la cual se construye el marco humanocéntrico. Lo que cambia es la lógica de aplicación: la tecnología deja de ser el fin para convertirse en el medio que facilita un entorno de trabajo más inclusivo y sostenible. Este fenómeno ha sido denominado por algunos expertos como Thinking 5.0, una mentalidad que anima a las organizaciones a repensar el propósito de la industria y a alinear la estrategia corporativa con los valores éticos y ambientales.

5.2 Los Tres Pilares de la Industria 5.0

La arquitectura estratégica de la Industria 5.0 se sustenta en tres pilares interconectados que redefinen la competitividad industrial en el siglo XXI: la centricidad humana, la sostenibilidad y la resiliencia. Estos pilares no son meros objetivos aislados, sino que constituyen un marco de viabilidad que permite a las empresas navegar por la complejidad geopolítica y económica actuales (Belkadi & Bachiri, 2025).

Centricidad Humana: El Trabajador como Activo y no como Recurso

La centricidad humana representa un cambio fundamental desde una visión en la que el trabajador es una pieza intercambiable de la maquinaria hacia otra en la que es el principal beneficiario y motor de la organización. Este pilar se centra en promover el talento, la diversidad y el empoderamiento, asegurando que el entorno de trabajo sea seguro, saludable y estimulante. La tecnología humanocéntrica no se diseña para que el humano se adapte a ella, sino para que la herramienta responda a las necesidades humanas, considerando factores como la carga cognitiva, la salud mental y la ergonomía avanzada.

La implementación de este pilar implica la creación de sistemas de Aumento Humano, en los que la inteligencia artificial explicable (XAI) y la robótica colaborativa

amplifican la intuición y la creatividad del operario. En lugar de competir con la máquina, el trabajador entra en una fase de coevolución, donde su juicio ético y su capacidad para resolver problemas contextuales se vuelven más valiosos a medida que las tareas rutinarias se automatizan.

Sostenibilidad: Respetando los Límites Planetarios

El pilar de la sostenibilidad exige que la industria lidere la transición verde y digital simultáneamente. Bajo la Industria 5.0, la sostenibilidad deja de ser una carga regulatoria para convertirse en una ventaja competitiva basada en la eficiencia de los recursos y la responsabilidad social corporativa. La adopción de principios de economía circular —reducción, reutilización y reciclaje— se facilita mediante el uso de gemelos digitales y sensores IoT que monitorean en tiempo real el consumo de energía y la generación de residuos (Shafique et al., 2024).

La industria se posiciona como proveedora de soluciones ante desafíos sociales, como el cambio climático y la preservación de los recursos. Esto incluye el diseño de productos con mayor longevidad, la integración de energías renovables en las plantas de producción y la optimización de las cadenas de suministro para minimizar la huella de carbono. La sostenibilidad en este paradigma es proactiva y busca generar un impacto netamente positivo en el ecosistema, en lugar de simplemente mitigar los daños.

Resiliencia: la ciencia de la adaptabilidad

La resiliencia en la Industria 5.0 se refiere a la capacidad de las organizaciones para anticipar, reaccionar y aprender de las crisis de manera sistemática y ágil. Frente a un entorno global caracterizado por la volatilidad, la resiliencia industrial se construye sobre tecnologías flexibles y una fuerza laboral capacitada para la toma de decisiones descentralizada. La capacidad de respuesta ante interrupciones, ya sean

desastres naturales o cambios bruscos en el mercado, se convierte en el indicador crítico de éxito.

Este pilar también abarca la resiliencia social, asegurando que las comunidades donde operan las industrias permanezcan estables y prósperas a pesar de los cambios tecnológicos. La resiliencia se manifiesta en la capacidad de reconfiguración de la cadena de suministro, lo que permite que las fábricas cambien sus líneas de producción rápidamente para responder a necesidades urgentes, como se observó durante la producción de ventiladores y equipos de protección en la fase inicial de la pandemia.

Tecnologías Habilitadoras y su Integración Socio-Técnica

La Industria 5.0 aprovecha las innovaciones tecnológicas para fomentar un entorno de producción que combina la precisión de las máquinas con el ingenio humano. La integración de estas tecnologías requiere un enfoque holístico que considere las implicaciones éticas, legales y sociales (ELSI).

IA Explicable (XAI) y Sistemas de Inteligencia Colaborativa

La inteligencia artificial es el cerebro de la transformación industrial, permitiendo funcionalidades como el autodiagnóstico y la predicción de mantenimiento. Sin embargo, la Industria 5.0 exige que esta IA sea explicable (XAI). La XAI aborda la opacidad de los algoritmos de caja negra, permitiendo que los operarios humanos comprendan las razones detrás de una recomendación o decisión tomada por la máquina (Ahangar et al., 2025).

Esta transparencia es vital para generar confianza en aplicaciones críticas, donde una decisión errónea puede tener impactos económicos o de seguridad significativos. Al facilitar la comunicación bidireccional entre humanos y máquinas,

la XAI permite una toma de decisiones informada y ética, asegurando que la supervisión humana siga siendo el pilar central de los procesos autónomos.

Robótica Colaborativa (Cobots) y Aumento Físico

Los robots colaborativos, o cobots, son el ejemplo más tangible de la sinergia humano-máquina. A diferencia de los robots industriales tradicionales, los cobots operan de forma segura en las proximidades de los humanos, asistiendo en tareas que requieren fuerza física o precisión repetitiva. Esta colaboración permite que el operario se centre en la personalización del producto y en el control de calidad, tareas que requieren sensibilidad y juicio humano (ver Tabla 14).

El aumento físico se extiende mediante el uso de:

1. **Exoesqueletos industriales:** dispositivos que proporcionan soporte biomecánico para reducir la carga en la espalda y las extremidades de los trabajadores, previniendo lesiones crónicas y extendiendo la vida laboral productiva, especialmente en una fuerza laboral que envejece.
2. **Sensores vestibles (Wearables):** Dispositivos que monitorean en tiempo real parámetros de salud y seguridad y alertan sobre riesgos ambientales o sobre la fatiga excesiva del trabajador.

El Gemelo Digital Humano (Human Digital Twin - HDT)

Mientras que los gemelos digitales tradicionales replican activos físicos, el concepto de Gemelo Digital Humano (HDT) crea representaciones virtuales de los trabajadores que incluyen sus capacidades, necesidades ergonómicas y estados fisiológicos. Los HDT permiten simular cómo interactuarán los trabajadores con nuevos entornos de producción o con máquinas antes de su implementación física, optimizando la seguridad y la eficiencia desde la fase de diseño.

Estos modelos virtuales capturan datos mediante dispositivos IIoT para evaluar el estrés, la carga cognitiva y el bienestar general. La aplicación de HDTs facilita una personalización extrema del entorno de trabajo, adaptando las estaciones de montaje a las características físicas e intelectuales individuales de cada empleado, lo que se traduce en una mayor satisfacción laboral y una reducción de los errores humanos.

Tabla 14: Tecnologías emergentes y su rol futuro

Tecnología	**Aplicación en Industria 5.0**	**Impacto en la Sostenibilidad/Resiliencia**
Blockchain	Trazabilidad total en la cadena de suministro.	Transparencia ética y reducción de fraude en materiales.
Computación Cuántica	Optimización de la logística y nuevos materiales.	Reducción drástica del consumo de energía de la computación.
Fabricación Aditiva (3D)	Producción local y personalizada bajo demanda.	Reducción masiva de residuos y de la huella de transporte.
Realidad Extendida (XR)	Formación inmersiva y asistencia remota.	Reducción de costos de viaje y mejora de la retención de habilidades.

Edge Computing	Procesamiento de datos en tiempo real en la planta.	Mayor resiliencia operativa ante fallos en la conectividad central.

5.3 Transformación del Talento Humano: Upskilling y Bienestar

La transición hacia la Industria 5.0 impone una presión sin precedentes sobre la gestión del talento. Se estima que para el año 2030, una parte significativa de la fuerza laboral mundial necesitará adquirir nuevas competencias para colaborar eficazmente con sistemas inteligentes.

El Desafío de las Habilidades: Hacia el Trabajador 5.0

La brecha de habilidades es uno de los mayores obstáculos para la adopción de este paradigma. La Industria 5.0 no solo requiere competencias técnicas en IA, análisis de datos y robótica, sino también habilidades transversales, o soft skills, que las máquinas no pueden replicar fácilmente. El pensamiento crítico, la creatividad, la inteligencia emocional y el liderazgo colaborativo se convierten en los pilares de la empleabilidad futura (Reyes et al., 2025).

Para cerrar esta brecha, las organizaciones y las instituciones educativas deben implementar programas de:

- **Upskilling:** Mejora de las habilidades actuales de los trabajadores para manejar tecnologías de aumento.
- **Reskilling:** Reciclaje profesional completo para trabajadores cuyos roles actuales serán desplazados por la automatización de bajo valor.

Gestión de la Salud Mental y Ergonomía Cognitiva

A diferencia de las revoluciones industriales anteriores, centradas en la salud

física, la Industria 5.0 otorga prioridad estratégica a la salud mental y a la ergonomía cognitiva. El ritmo acelerado de la digitalización y el monitoreo constante pueden generar ansiedad, estrés y fatiga mental. Las tecnologías de la Industria 5.0 se utilizan para mitigar estos riesgos mediante el desarrollo de entornos de trabajo empáticos.

Existen investigaciones que proponen el uso de IA y el procesamiento del lenguaje natural (NLP) para analizar emociones en conversaciones escritas o en expresiones faciales, lo que permite identificar signos tempranos de agotamiento (*burnout*). El monitoreo de la variabilidad de la frecuencia cardíaca (HRV) mediante relojes inteligentes permite a las empresas implementar intervenciones preventivas, como sugerir pausas o ajustes en la carga de trabajo, antes de que se produzcan incidentes de salud.

El Envejecimiento de la Fuerza Laboral y la Inclusión

Muchas economías industrializadas enfrentan el desafío de una fuerza laboral envejecida. La Industria 5.0 aborda este problema mediante tecnologías de asistencia que compensan el declive de las capacidades físicas y cognitivas, permitiendo que los trabajadores experimentados permanezcan activos y transmitan su conocimiento crítico a las nuevas generaciones. La inclusión también abarca a personas con discapacidades, mediante interfaces adaptativas y robótica de apoyo, para garantizar que todos puedan participar en la creación de valor. La adopción de la Industria 5.0 no es uniforme y está influida por las prioridades nacionales y regionales, lo que refleja visiones distintas pero convergentes sobre el futuro de la sociedad.

La Unión Europea: El Plan de Transición 5.0

Europa ha liderado la formalización de la Industria 5.0 como parte de su estrategia para alcanzar la neutralidad climática y la soberanía tecnológica. El Plan de Transición 5.0, lanzado en 2024, destina aproximadamente 13.000 millones de euros

para apoyar a las empresas en su transformación digital y energética. Este plan se estructura en torno a créditos fiscales que incentivan la reducción del consumo de energía y la adopción de tecnologías limpias.

La política europea enfatiza que la competitividad a largo plazo depende de la inversión en habilidades y de la creación de una industria atractiva para las nuevas generaciones, como la Generación Z, que valora el propósito, la flexibilidad y la sostenibilidad en su carrera profesional.

Japón: La Visión de la Sociedad 5.0

Japón fue pionero en este concepto con su propuesta de la Sociedad 5.0 en 2016. Mientras que la Industria 5.0 nace del sector manufacturero, la Sociedad 5.0 es una visión transversal que busca resolver problemas sociales (como el envejecimiento de la población y la falta de competitividad) mediante la integración masiva del ciberespacio y del espacio físico. El enfoque japonés utiliza la robótica y la IA no solo para la producción, sino también para el cuidado de ancianos, el transporte autónomo en zonas rurales y la creación de ciudades inteligentes (*Smart Cities*) que mejoren la calidad de vida general de los ciudadanos.

5.4 Desafíos Éticos, Legales y la Brecha Digital en las PYMES

A pesar de las promesas de bienestar y sostenibilidad, la implementación de la Industria 5.0 enfrenta desafíos estructurales que podrían exacerbar las desigualdades existentes.

El Dilema Ético de la Vigilancia y el Consentimiento

La recopilación intensiva de datos biométricos y de comportamiento para alimentar los HDTs y los sistemas de salud mental plantea serias preocupaciones sobre la privacidad y la vigilancia en el lugar de trabajo. Existe el riesgo de que las

tecnologías diseñadas para el bienestar se utilicen para la microgestión o la discriminación basada en el rendimiento físico o cognitivo.

Los marcos éticos, como la serie de estándares IEEE P7000, buscan mitigar estos riesgos mediante procesos de diseño basados en valores. Estos estándares recomiendan:

- **Transparencia:** Los algoritmos deben ser auditables y su lógica debe ser comprensible para los usuarios.
- **Privacidad por diseño:** Los datos deben minimizarse y protegerse contra el uso indebido desde la concepción del sistema.
- **Agencia Humana:** Los sistemas nunca deben delegar decisiones críticas de vida o muerte a máquinas sin una supervisión humana efectiva.

La Brecha Digital en las Pequeñas y Medianas Empresas (PYMES)

Las PYMES constituyen la columna vertebral de la economía global, representando el 99% de las empresas en la UE y generando entre el 60% y el 70% del empleo mundial. Sin embargo, la transición a la Industria 5.0 resulta especialmente difícil para ellas debido a la escasez de recursos financieros, la falta de personal calificado y el alto costo de la infraestructura tecnológica.

Muchas PYMES todavía operan con tecnologías más alineadas con la Industria 3.0, lo que las coloca en riesgo de extinción si no logran adaptarse a las nuevas exigencias de las cadenas de suministro globales. Para sobrevivir, se recomienda que estas empresas se centren en innovaciones incrementales y busquen asociaciones con centros de investigación y proveedores de servicios de seguridad gestionados (MSSP) para mitigar los riesgos de ciberseguridad (Mendizábal et al., 2019).

Modelos de Madurez y la Ruta de Implementación

Para transitar hacia la Industria 5.0, las empresas necesitan un marco de evaluación que les permita identificar su estado actual y priorizar inversiones. Los modelos de madurez específicos para la Industria 5.0 están empezando a surgir, integrando métricas de sostenibilidad y centricidad humana junto con las de la digitalización tradicional.

Etapas de la Transformación Organizacional

La literatura académica y los informes de expertos sugieren una ruta de adopción modular y en fases, en lugar de un enfoque de cambio disruptivo de gran magnitud.

1. **Evaluación de la Madurez Digital y Cultural:** Antes de introducir tecnologías avanzadas, la organización debe evaluar si su fuerza laboral cuenta con alfabetización digital básica y si existe una cultura de apertura a la experimentación.
2. **Alineación Estratégica con los 3 Pilares:** Definir objetivos claros relacionados con la reducción de la huella de carbono, la mejora de la ergonomía o el fortalecimiento de la resiliencia ante crisis de suministro.
3. **Implementación de Proyectos Piloto (Cobots/IA):** Introducir tecnologías de colaboración en procesos específicos de bajo riesgo para demostrar su valor y generar confianza entre los trabajadores.
4. **Desarrollo de un Ecosistema de Aprendizaje Continuo:** Establecer canales de retroalimentación en los que los operarios participen en el diseño de soluciones tecnológicas, asegurando que las herramientas realmente resuelvan sus problemas diarios.
5. **Escalamiento y Optimización mediante Datos:** Utilizar la analítica avanzada para refinar los procesos, buscando el equilibrio óptimo entre la productividad

de la máquina y el bienestar humano.

Análisis de Casos: La Industria 5.0 en la Práctica

Varias empresas globales han comenzado a reportar beneficios tangibles al aplicar los principios de la Industria 5.0, lo que valida el concepto mediante resultados concretos. Mediante su programa *Digital Factory*, Siemens ha utilizado gemelos digitales para reducir el consumo de energía en sus plantas en más del 20%. La integración de IA para la gestión inteligente de la energía ha permitido a la empresa adaptar su producción en tiempo real a la disponibilidad de las fuentes renovables, mejorando tanto su rentabilidad como su perfil de sostenibilidad (Pérez et al., 2024).

Ferrero en Italia utiliza gemelos digitales para simular escenarios logísticos complejos. Al optimizar las rutas y la carga de los vehículos basándose en datos en tiempo real, han logrado mejorar los tiempos de entrega y reducir significativamente el desperdicio de productos y de combustible, alineándose con los objetivos de resiliencia y respeto al medio ambiente.

La Industria 5.0 marca el fin de la era en la que la tecnología se desarrollaba de forma aislada de sus consecuencias sociales y ambientales. El giro humanocéntrico no es simplemente un imperativo ético, sino una necesidad pragmática para la supervivencia de la industria en un mundo volátil y con recursos limitados. Al posicionar al trabajador como un activo estratégico y a la sostenibilidad como un motor de innovación, las empresas no solo protegen el futuro del planeta, sino que también aseguran su propia resiliencia y competitividad a largo plazo.

La transición exitosa requiere un compromiso conjunto de líderes industriales, responsables políticos y trabajadores. Para las empresas, el camino implica invertir tanto en la capacitación de la fuerza de trabajo como en la infraestructura digital. Para los gobiernos, el desafío radica en crear marcos regulatorios e incentivos financieros

que mitiguen los riesgos para las PYMES y aseguren que los beneficios de la transformación se distribuyan equitativamente en la sociedad.

En última instancia, la Industria 5.0 representa una reconciliación entre la eficiencia de la máquina y la creatividad del alma humana. Al trabajar en perfecta simbiosis, esta nueva era promete no solo una producción más inteligente, sino también una sociedad más justa, resiliente y en armonía con los límites naturales de nuestro mundo.

Capítulo VI
Integración neuro-simbólica

La evolución de la inteligencia artificial ha estado históricamente marcada por una oscilación pendular entre dos filosofías divergentes: el enfoque simbólico, basado en la manipulación de reglas lógicas y representaciones explícitas, y el enfoque conexionista o subsimbólico, que confía en el aprendizaje estadístico a partir de grandes volúmenes de datos. En la actualidad, nos encontramos en lo que la comunidad científica denomina el tercer verano de la IA, un periodo de avances sin precedentes impulsado principalmente por el éxito de las redes neuronales profundas y los modelos de lenguaje de gran escala (LLM). Sin embargo, a medida que estos sistemas puramente estadísticos se enfrentan a desafíos críticos en términos de explicabilidad, robustez lógica y eficiencia en el uso de datos, surge la integración neuro-simbólica como el paradigma definitivo para alcanzar una inteligencia artificial más general, fiable y capaz de razonar como el ser humano.

La tesis central que define el estado actual de la investigación sostiene que el futuro de la IA no reside en la victoria de una de estas facciones, sino en su síntesis técnica y conceptual. Esta integración busca amalgamar la capacidad de percepción y reconocimiento de patrones de las redes neuronales con la estructura lógica, la transparencia y la inferencia multihop del razonamiento simbólico. El objetivo es transitar desde sistemas que simplemente imitan la inteligencia mediante asociaciones estadísticas hacia arquitecturas que construyen activamente modelos del mundo, manejan la causalidad y respetan restricciones formales en tiempo real.

6.1 Marco teórico y taxonomía de la integración

La inteligencia artificial neuro-simbólica (NeSy) se define formalmente como

un marco que fusiona los dominios de la IA simbólica y de las redes neuronales para crear modelos híbridos superiores. Este movimiento se inspira profundamente en la psicología cognitiva, en particular en la teoría de los dos sistemas de Daniel Kahneman (Villalobos et al., 2025). Mientras que el Sistema 1 es rápido, intuitivo, paralelo y capaz de procesar señales sensoriales complejas —análogo a las redes neuronales—, el Sistema 2 es lento, deliberado, secuencial y basado en reglas —representado por el razonamiento simbólico—. La integración neuro-simbólica pretende replicar esta sinergia en sistemas computacionales para superar las cajas negras del aprendizaje profundo.

Para sistematizar el estudio de este campo en rápida expansión, investigadores como Colelough y Regli han propuesto una taxonomía basada en cinco áreas fundamentales de investigación, cuya distribución de esfuerzos revela las prioridades actuales de la comunidad científica (ver Tabla 15).

Tabla 15: Taxonomía basada en cinco áreas fundamentales de investigación

Área de Investigación	**Descripción de la Integración**	**Distribución de Esfuerzos (2020-2024)**
Aprendizaje e Inferencia	Combinación de procesos de aprendizaje y de razonamiento, mediante lógica diferenciable y razonamiento dinámico multisectorial.	63%
Representación del	Integración de representaciones simbólicas y neuronales;	44%

Conocimiento	desarrollo de grafos de conocimiento de sentido común.	
Lógica y Razonamiento	Incorporación de métodos basados en la lógica (booleana, difusa o probabilística) en las arquitecturas neuronales.	35%
Explicabilidad y Confianza	Desarrollo de modelos interpretables que permitan auditorías y verifiquen la fiabilidad de las decisiones.	28%
Metacognición	Sistemas capaces de monitorear, evaluar y ajustar de forma autónoma sus propios procesos de razonamiento.	5%

Esta distribución estadística evidencia que, aunque existe un sólido cuerpo de trabajo en aprendizaje e inferencia, persiste una brecha crítica en áreas como la metacognición y la confianza, elementos vitales para el despliegue de la IA en entornos de alto riesgo y para la consecución de una autonomía adaptativa real.

Clasificaciones arquitectónicas según el grado de acoplamiento

1. **Arquitecturas en Cascada o Pipeline:** El componente neuronal actúa como un preprocesador sensorial que transforma datos no estructurados (imágenes, audio) en símbolos sobre los cuales opera un razonador lógico tradicional.
2. **Arquitecturas Anidadas:** Un algoritmo simbólico tradicional sirve como marco

computacional principal, pero llama a una red neuronal como una subrutina especializada para tareas específicas que resultan difíciles de codificar mediante reglas explícitas, como el reconocimiento visual en sistemas de navegación.

3. **Integración Directa mediante Funciones de Pérdida (Semantic Loss):** Se utilizan reglas simbólicas para regularizar el entrenamiento de la red neuronal. La lógica no reside en la arquitectura, sino que actúa como una restricción durante el aprendizaje para asegurar que las predicciones de la red respeten leyes físicas o normativas legales.
4. **Compilación Lógica en Arquitectura (Wired/Monolithic):** El conocimiento simbólico se compila directamente en la estructura de la red neuronal. En estos sistemas, cada neurona o capa tiene una interpretación lógica definida, lo que garantiza la transparencia desde el diseño.
5. **Representaciones Tensoriales de Lógica (Monolithic Tensor):** Se utilizan tensores y operaciones diferenciables para representar funciones lógicas, lo que permite un entrenamiento de extremo a extremo mientras se mantiene la estructura de las fórmulas de primer orden.

Ecosistemas de razonamiento y marcos de trabajo destacados

El desarrollo de la IA neurosimbólica ha dado lugar a marcos de trabajo robustos que ya demuestran su eficacia en tareas complejas de razonamiento y percepción. Estos proyectos no solo son teóricos, sino que también incluyen implementaciones de software que permiten a los investigadores razonar mientras aprenden y aprender mientras razonan.

Redes Neuronales Lógicas (Logical Neural Networks - LNN)

Las Redes Neuronales Lógicas, desarrolladas por IBM Research, representan uno de los intentos más rigurosos de unificar el aprendizaje neuronal con la lógica formal. En una LNN, cada neurona tiene una correspondencia exacta, uno a uno, con

un componente de una fórmula lógica en un sistema de lógica real ponderada. A diferencia de las redes convencionales, donde los pesos son opacos, las LNN mantienen una interpretabilidad total: el grafo de la red refleja directamente la estructura de las fórmulas lógicas que representa (Sarmiento, 2020).

Las innovaciones clave de las LNN incluyen:

- **Funciones de activación restringidas:** Implementan funciones de verdad de operadores lógicos como And, Or, Not, Implies y cuantificadores de lógica de primer orden como Forall y Exists.
- **Inferencia mediante límites (Bounds):** En lugar de valores escalares simples, las LNN operan con límites sobre los valores de verdad. Esto permite al sistema expresar estados de conocimiento parciales, desconocidos o contradictorios.
- **Inferencia bidireccional:** La red puede procesar información tanto hacia arriba (deducción) como hacia abajo (abducción), lo que permite, por ejemplo, probar una premisa si se conoce la conclusión y el resto del argumento.
- **Resiliencia a la inconsistencia:** El entrenamiento minimiza una función de pérdida que captura las contradicciones lógicas, lo que hace que el sistema sea robusto frente a bases de conocimiento con información contradictoria.

6.2 DeepProbLog: Integración de lógica probabilística y aprendizaje profundo

DeepProbLog es un lenguaje de programación lógica probabilística que extiende el marco de ProbLog mediante la introducción de predicados neuronales. Este sistema permite que las probabilidades de los hechos lógicos sean determinadas por redes neuronales que procesan datos sensoriales, creando una interfaz fluida entre la percepción de bajo nivel y el razonamiento de alto nivel.

La arquitectura de DeepProbLog permite el aprendizaje de extremo a extremo

a partir de ejemplos, incluso cuando los predicados neuronales internos no están etiquetados explícitamente. Por ejemplo, en una tarea de suma de dígitos manuscritos, el sistema solo necesita el resultado final de la suma para entrenar las redes encargadas de reconocer los dígitos individuales, utilizando el conocimiento previo de las reglas de la aritmética para guiar el proceso de aprendizaje (Iovane & Iovane, 2026). Este enfoque ha demostrado ser significativamente más eficiente en el uso de datos que los métodos puramente neuronales, lo que permite una generalización superior a tareas más complejas sin necesidad de un reentrenamiento extensivo.

El Aprendiz de Conceptos Neuro-Simbólico (NS-CL)

El modelo NS-CL, desarrollado por investigadores del MIT y de Stanford, propone un paradigma en el que la percepción visual y la comprensión del lenguaje se aprenden de forma conjunta y recíproca. El sistema consta de tres módulos principales: un módulo de percepción que extrae representaciones centradas en objetos, un analizador semántico que traduce preguntas de lenguaje natural en programas simbólicos ejecutables, y un ejecutor de programas cuasi-simbólicos que opera sobre las representaciones visuales.

Una de las características más potentes del NS-CL es su capacidad de aprendizaje continuo y de generalización compositiva. Al representar el conocimiento como conceptos modulares y combinables, el sistema puede responder preguntas complejas sobre escenas que contienen objetos o combinaciones de atributos que nunca vio durante el entrenamiento (Liang et al., 2025). Además, el uso de un aprendizaje por currículo permite que el sistema navegue eficazmente por el vasto espacio compositivo del lenguaje y de la visión, comenzando con conceptos básicos y escalando hacia razonamientos abstractos.

Superación de las limitaciones de los modelos de lenguaje de gran escala (LLM)

El auge de los LLM ha generado un debate sobre si el escalado masivo de datos y parámetros es suficiente para alcanzar la inteligencia artificial general (AGI). Los críticos sostienen que los LLM son loros estocásticos que predicen el siguiente token a partir de correlaciones estadísticas, pero carecen de una comprensión real del contexto, de la causalidad y del mundo físico (Scotto, 2025). En este sentido, la IA neurosimbólica se presenta como el camino necesario para dotar a las máquinas de un mundo-modelo sólido y de un razonamiento fiable (ver Tabla 16).

Tabla 16: Limitación de los LLM

Limitación de los LLM puros	**Solución Neuro-Simbólica**	**Impacto esperado**
Alucinaciones y falsedades	Integración de reglas explícitas y de bases de conocimiento verificadas.	Reducción drástica de la información generada en entornos críticos.
Opacidad (Caja Negra)	Explicaciones lógicas, paso a paso, de cada decisión.	Cumplimiento de auditorías legales y normativas, como el GDPR.
Ineficiencia de datos	Uso de conocimiento previo para aprender con menos ejemplos.	Reducción del 50% al 75% en los tiempos de migración y de entrenamiento.

Fragilidad lógica	Aplicación del razonamiento formal y de las leyes de causalidad.	Capacidad para realizar razonamientos de varios pasos sin errores triviales.
Falta de sentido común	Inclusión de grafos de conocimiento de sentido común (ConceptNet).	Comportamiento más alineado con las expectativas humanas intuitivas.

La IA neuro-simbólica aborda directamente el problema de las alucinaciones al embeber reglas lógicas que restringen la salida de los modelos neuronales, asegurando que los resultados sean factualmente precisos y lógicamente consistentes. El consenso de los expertos sugiere un cambio de modelos monolíticos gigantescos a ecosistemas modulares, donde la IA simbólica actúa como el centro consciente de procesamiento lento y deliberado, mientras que los componentes neuronales gestionan la percepción rápida (Parellada & Rovira, 2025).

6.3 Aplicaciones de alto impacto en sectores regulados

La necesidad de una inteligencia confiable y explicable es particularmente aguda en sectores en los que las decisiones tienen consecuencias legales, financieras o de seguridad personal. La IA neuro-simbólica está demostrando ser la solución ideal para estos entornos de alto riesgo.

Finanzas, auditoría y cumplimiento regulatorio

En el sector financiero, la transparencia y la rendición de cuentas son requisitos ineludibles. Los modelos neuro-simbólicos permiten no solo detectar patrones de fraude complejos, sino también justificar ante los reguladores por qué se ha marcado

una transacción específica.

Salud, medicina de precisión y biotecnología

En medicina, la IA neurosimbólica facilita la creación de sistemas capaces de razonar sobre guías clínicas y la literatura médica para ofrecer diagnósticos y recomendaciones de tratamiento personalizadas. A diferencia de los modelos neuronales puros, que pueden emitir diagnósticos basados en correlaciones espurias, los sistemas NeSy pueden rastrear sus conclusiones hasta la evidencia biológica o médica comprobable. El uso de arquitecturas neurosimbólicas en el descubrimiento de fármacos permite no solo identificar candidatos prometedores, sino también proporcionar explicaciones sobre por qué se espera que una molécula interactúe con un objetivo biológico específico. Esto reduce drásticamente los costos y los tiempos de desarrollo, lo que permite una validación experimental más dirigida y eficiente.

Seguridad crítica y verificación industrial

La verificación de cumplimiento en sistemas industriales, como paneles de control eléctrico, es una tarea que requiere una precisión absoluta y el estricto respeto a las normativas de seguridad. La IA neuro-simbólica permite procesar simultáneamente esquemas eléctricos, fotografías de instalaciones y diagramas de cableado. Mientras el componente neuronal identifica objetos y conexiones, el componente simbólico razona sobre las reglas de seguridad y los protocolos, detectando problemas de cumplimiento en minutos en lugar de horas.

En el ámbito marítimo y de aviación, la integración de datos de sensores en tiempo real con reglas de seguridad predefinidas asegura que las decisiones autónomas estén fundamentadas lógicamente, minimizando el riesgo de maniobras peligrosas derivadas de interpretaciones erróneas de los datos estadísticos. Si un sistema recomienda cambiar la ruta de un vuelo debido a una anomalía, el

componente simbólico puede explicar la causa de forma humana y facilitar la colaboración entre pilotos y máquinas.

6.4 Desafíos sistémicos: hardware, escalabilidad y brechas de investigación

A pesar de su innegable potencial, la adopción masiva de la IA neurosimbólica enfrenta obstáculos técnicos significativos, especialmente en lo que respecta a la infraestructura de cómputo y a la estandarización de la investigación.

La divergencia entre algoritmos y hardware

La trayectoria actual del hardware de IA está optimizada casi exclusivamente para el aprendizaje profundo, centrándose en la multiplicación de matrices y en convoluciones masivamente paralelas en GPUs. Por el contrario, las operaciones simbólicas y lógicas suelen ser memory-bound (limitadas por el ancho de banda de memoria) y presentan un flujo de control complejo y accesos irregulares a los datos.

Esta ineficiencia del hardware para procesar operaciones simbólicas genera cuellos de botella que dificultan la escalabilidad de estos sistemas. Para superar esto, los investigadores sugieren soluciones de optimización en las capas de software y de hardware, incluyendo el desarrollo de aceleradores específicos para la lógica neuro-simbólica y de arquitecturas de vectores simbólicos (VSA).

El vacío en metacognición y autorregulación

Una de las revelaciones más impactantes de las revisiones sistemáticas de 2024 es la escasez de investigación sobre la metacognición en el campo de la IA. La metacognición, definida como la capacidad de pensar sobre el propio pensamiento, es crucial para que los sistemas de IA puedan monitorear su rendimiento, detectar cuándo su razonamiento es erróneo y ajustar sus procesos de aprendizaje de forma

autónoma.

El descuido en esta área limita la autonomía y la fiabilidad de los sistemas en entornos dinámicos, donde la capacidad de corregir errores en tiempo real resulta fundamental para mantener la confianza del usuario. La integración de marcos de metacognición permitiría que la IA actuara como un controlador superior que dirigiera eficazmente los recursos del sistema hacia la subtarea correcta, mejorando la toma de decisiones complejas y la resolución de problemas.

Hacia el horizonte de 2026: Agentic AI y el camino a la AGI

Hoy en día, se espera que la IA neurosimbólica alcance un punto de madurez comercial en el que la discusión pase de si la IA puede realizar una tarea a si puede razonar de forma responsable y justificar sus decisiones. Esta evolución está intrínsecamente ligada al surgimiento de la Agentic AI, ecosistemas de agentes de razonamiento que colaboran entre sí, cada uno con roles explícitos y pistas de auditoría claras.

El marco G-I-A y la soberanía del razonamiento

La evaluación de los futuros sistemas neuro-simbólicos se centrará en el marco G-I-A (Grounding, Instructibility, Alignment).

- **Grounding (Fundamentación):** La capacidad de vincular el reconocimiento de patrones con conceptos del mundo real y con restricciones físicas.
- **Instructibility (Instructibilidad):** La posibilidad de que los sistemas adapten su comportamiento en respuesta al feedback de expertos humanos sin necesidad de un reentrenamiento masivo.
- **Alignment (Alineación):** Asegurar que las decisiones de la IA no solo sean estadísticamente precisas, sino que también cumplan con los objetivos organizacionales, legales y éticos.

La integración del razonamiento causal se considera el avance más transformador de esta era, al permitir estrategias de defensa cibersegura proactivas y una comprensión profunda de las relaciones causa-efecto que trascienden la mera correlación. Sin embargo, los expertos también advierten sobre las implicaciones del doble uso de estos sistemas, ya que la autonomía avanzada podría utilizarse para orquestar ciberataques más sofisticados y adaptativos.

La integración neuro-simbólica representa mucho más que una tendencia tecnológica pasajera; es la base necesaria para la próxima generación de inteligencia artificial a escala empresarial y social. Al fusionar la adaptabilidad y el poder de aprendizaje de las redes neuronales con el rigor y la transparencia de la lógica simbólica, estamos construyendo sistemas que no solo imitan la fluidez humana, sino que también heredan nuestra capacidad de juicio y responsabilidad (Villalobos et al., 2025).

El éxito de este paradigma dependerá de la colaboración interdisciplinaria entre neurocientíficos, lógicos, ingenieros de hardware y expertos en ética. A medida que los costos de entrenamiento de los modelos puramente neuronales se vuelven insostenibles y la presión regulatoria aumenta, la IA neuro-simbólica ofrece una alternativa eficiente, explicable y robusta. En última instancia, la transición de una IA que reconoce patrones a otra que comprende significados marcará el inicio de una nueva era en la interacción entre humanos y máquinas, donde la tecnología actúa como un socio de decisión confiable y no simplemente como una herramienta estadística opaca.

Conclusión

A pesar de los avances, este libro dedica secciones críticas a los desafíos que aún persisten en 2026. La brecha de habilidades es quizás el mayor obstáculo para la implementación exitosa de la inteligencia artificial; mientras la tecnología avanza a un ritmo exponencial, la alfabetización digital de la fuerza laboral y de los tomadores de decisiones a menudo se queda atrás. El 62% de las organizaciones reportan carecer de las habilidades necesarias para gestionar datos de IA, lo que genera riesgos de seguridad e ineficiencias operativas.

La ética de la IA también se enfrenta a dilemas complejos relacionados con la privacidad y la propiedad intelectual. La formación de modelos de IA con datos protegidos por derechos de autor sigue siendo un terreno de disputas legales intensas, al igual que el uso de deepfakes para la desinformación política y el acoso personal. La responsabilidad en la toma de decisiones automatizada —quién es responsable cuando un agente de IA causa un daño— es una pregunta que los marcos legales están empezando a responder, pero que aún requiere una deliberación profunda sobre la agencia y la personalidad jurídica de los sistemas autónomos.

Hoy estamos presenciando el surgimiento de la inteligencia ecosistémica, donde la IA ya no es un sistema aislado, sino el pegamento que mantiene unidos los mundos físico y digital. Las ciudades inteligentes, las cadenas de suministro resilientes y las redes de salud adaptativas operan como organismos vivos en los que los agentes inteligentes colaboran entre sí y con

los humanos para resolver problemas en tiempo real.

Este libro ha sido diseñado para guiar al lector a través de este laberinto de innovaciones y desafíos técnicos en torno al aprendizaje profundo, las arquitecturas neurosimbólicas y los marcos de gobernanza y, en síntesis, proporcionar las herramientas conceptuales y técnicas para construir sistemas poderosos, confiables, justos y profundamente humanos. Pues, este es el ciclo en el que la humanidad decidió que la tecnología debe servir para elevar nuestra capacidad de razonamiento, creatividad y compasión.

En conclusión, la transición de la magia algorítmica a la ingeniería de sistemas confiables ha comenzado. Aquellos que actúen ahora para alinear sus desarrollos con estos estándares de responsabilidad no solo mitigarán riesgos, sino que también ganarán una ventaja competitiva en un mundo donde la confianza es el activo más valioso de la era digital.

En América Latina, la transformación es igualmente profunda, aunque enfrenta desafíos de infraestructura y de financiamiento. Países como Costa Rica y Panamá están liderando la tropicalización del marco regulatorio europeo, adaptando las categorías de riesgo a las necesidades locales y promoviendo una IA que ayude a revitalizar las lenguas indígenas y a proteger la biodiversidad amazónica. Por lo tanto, la colaboración regional es clave, con redes como el AI Global Education Network (AIGEN) que facilitan la transferencia de conocimiento y el desarrollo de modelos de lenguaje adaptados a las variantes lingüísticas y culturales de la región.

Finalmente, la evolución técnica de la inteligencia artificial avanza en paralelo con una maduración sin precedentes de los marcos legales y éticos. Hemos pasado de la era de la autorregulación industrial a una fase de

aplicación estricta de leyes vinculantes. El enfoque regulatorio global ha convergido hacia un modelo basado en el riesgo, en el que la intensidad de las obligaciones legales es proporcional al impacto potencial del sistema de IA sobre los derechos fundamentales y la seguridad física de las personas.

Bibliografía

Ahangar, M.N., Farhat, Z.A., & Sivanathan, A. (2025). AI Trustworthiness in Manufacturing: Challenges, Toolkits, and the Path to Industry 5.0. *Sensors (Basel, Switzerland)*, *25*(14), 4357. https://doi.org/10.3390/s25144357

Amzil, K., Saidi, R. & Cherif, W. (2025). Mejora de procesos industriales críticos con modelos de inteligencia artificial. *Actas de Ingeniería* , *112* (1), 75. https://doi.org/10.3390/engproc2025112075

Aparicio, B. (2025). Directrices de la Unión Europea sobre prácticas prohibidas de Inteligencia Artificial. *Revista CESCO de Derecho de Consumo*. 53. https://doi.org/10.18239/RCDC_2025.53.3678

Araya Paz, C. (2021). Transparencia algorítmica ¿un problema normativo o tecnológico?. *CUHSO (Temuco)*, *31*(2), 306-334. https://dx.doi.org/10.7770/cuhso-v31n2-art2196

Arguedas Vega, D. (2024). Robótica y Cirugía: Innovaciones y Tendencias Actuales en la Práctica Quirúrgica. *Revista Veritas De Difusão Científica*, *5*(2), 464–478. https://doi.org/10.61616/rvdc.v5i2.97

Baleriola Escudero, E., Piñones Valenzuela, R., Rivera-Aguilera, G., Cáceres Serrano, P., & Tirado-Serrano, F.J. (2025). Tecnología y gubernamentalidad algorítmica: Transformaciones sociales y políticas en la Era de la Inteligencia Artificial. *Psicoperspectivas,* 24(2). https://www.dx.doi.org/10.5027/psicoperspectivas-vol24-issue1-fulltext-3547

Belkadi, A. & Bachiri, M. (2025). Hacia una Industria 5.0 Potenciada por la IA: Un Marco Teórico. *Actas de Ingeniería* , *112* (1), 2. https://doi.org/10.3390/engproc2025112002

Da Costa Pimenta, C.C., Zarzuelo Prieto, D., Balan Garcia, A., Goicochea Calderón, J.A., & Beltrán Sánchez, S. (2025). Industria 4.0: Impacto de la Digitalización y la Automatización en la Transformación Social e Industrial. *Revista Inclusiones*, *12*(1), 216–243. https://doi.org/10.58210/fprc3592

Enriquez, D.W., & Raraz, J.G. (2024). Uso de la inteligencia artificial en el laboratorio clínico y la diabetes, en países en vías de desarrollo: El control de la calidad y la estimación de la glucosa a partir de la hemoglobina glicada. (2024). *Revista Médica Basadrina*, *18*(1), 48-51. https://doi.org/10.33326/26176068.2024.1.1977

Escobar, A.L., & Giraldo Suárez, E. (2005). Implementación de algoritmos de recorrido de grafos para el cálculo de la regulación en redes de distribución radiales. *Scientia Et Technica.* 21(27), 33-35

Estrada Merino, A. (2025). Inteligencia artificial: hacia una gobernanza ética. *Acta Herediana*, *68*(2), 33–40. https://doi.org/10.20453/ah.v68i2.7505

Frankish, K., & Ramsey, W.M. (2014). *The Cambridge handbook of artificial intelligence.* Cambridge: Cambridge University Press

Garikapati, K., Torres Manrique, J.I., & Goyal, R. (2025). Contribución de la IA a la gobernanza de la sostenibilidad medioambiental en India y Perú: estudio comparativo. *Revista Oficial Del Poder Judicial*, *17*(23), 83-112.

https://doi.org/10.35292/ropj.v17i23.1138

Ge, M., Ohtani, K., Niu, Y., Zhang, Y., & Takeda, K. (2025). VLA-MP: A Vision-Language-Action Framework for Multimodal Perception and Physics-Constrained Action Generation in Autonomous Driving. *Sensors*, *25*(19), 6163. https://doi.org/10.3390/s25196163

Iovane, G. y Iovane, G. (2026). Sofimática y tiempo complejo 2D para mitigar las alucinaciones en los LLM para nuevos sistemas de información inteligentes en la transformación digital. *Applied Sciences* , *16* (1), 288. https://doi.org/10.3390/app16010288

Liang, B., Wang, Y. y Tong, C. (2025). Razonamiento de IA en la era del aprendizaje profundo: De la IA simbólica a la IA neuronal-simbólica. *Matemáticas* , *13* (11), 1707. https://doi.org/10.3390/math13111707

Linares, F. (2024). Inteligencia Artificial y Ciberdefensa. *Revista Seguridad y Poder Terrestre.* 3(4), 139-150. https://doi.org/10.56221/spt.v3i4.72

Llacsa Puma, L.J., Meleán Romero, R.A., Guadalupe-Zevallos, O.G., & Mamani Cachicatari, G. (2026). Marco legal para regular el uso de la inteligencia artificial en educación: Reflexiones sobre derechos fundamentales en edades tempranas. *Cuestiones Políticas*, *44*(84), 75-85. https://doi.org/10.5281/zenodo.18735101

McKinsey & Company. (2024). «What is artificial general intelligence (AGI)?,» 21 marzo 2024. [En línea]. Available: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-artificial-general-intelligence-agi

Mendizábal, G., Sánchez, A., & Kurkzyn, P. (2019). Industria 4.0. Trabajo y seguridad. Ciudad de México: Universidad Nacional Autónoma de México. https://archivos.juridicas.unam.mx/www/bjv/libros/12/5645/20.pdf

Naciones Unidas (UN). (2024). *Gobernanza de la Inteligencia Artificial en beneficio de la Humanidad: Informe final.* New York: United Nations Publications. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_es.pdf

OECD/CAF. (2022), *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe, Estudios de la OCDE sobre Gobernanza Pública*, OECD Publishing, Paris, https://doi.org/10.1787/5b189cb4-es

Okdem, S. & Okdem, S. (2024). Inteligencia artificial en ciberseguridad: Una revisión y un estudio de caso. *Applied Sciences*, *14* (22), 10487. https://doi.org/10.3390/app142210487

Parellada Guillamón, S., & Rovira Martorell, J. (2025). Gubernamentalidad neuroalgorítmica: Inteligencia Artificial y producción de neurosujetos. *Psicoperspectivas*, *24*(2), 1-19. https://dx.doi.org/10.5027/psicoperspectivas-vol24-issue2-fulltext-3460

Pérez-Domínguez, L.A. (2024). Las principales tecnologías de la era de la industria 5.0. *Revista Ingenio*, *21*(1), 60–70. https://doi.org/10.22463/2011642X.4352

Porcelli, A.M. (2020). La inteligencia artificial y la robótica: sus dilemas sociales, éticos y jurídicos. *Derecho global. Estudios sobre derecho y justicia*, *6*(16), 49-105.

https://doi.org/10.32870/dgedj.v6i16.286

Reyes, A.Y., Jiménez, C.P. & Poblano, E.R. (2025). *Aplicaciones de la Inteligencia Artificial en la Administración de las Organizaciones.* Jalisco: Editorial Centro de Estudios e Investigaciones para el Desarrollo Docente CENID. https://dialnet.unirioja.es/descarga/libro/1017142.pdf

Rodríguez-Martínez, E.A., Flores-Fuentes, W., Achakir, F., Sergiyenko, O. y Murrieta-Rico, F.N. (2025). Navegación y percepción basadas en visión para robots autónomos: Sensores, SLAM, estrategias de control y aplicaciones multidominio: Una revisión. *Eng*, *6* (7), 153. https://doi.org/10.3390/eng6070153

Russel, S.J., & Norvig, P. (2004). *Inteligencia artificial: un enfoque moderno.* Madrid: Pearson Educación. http://jdelagarza.fime.uanl.mx/IA/Libros/inteligencia-artificial-un-enfoque-moderno-stuart-j-russell.pdf

Sarmiento-Ramos, J. (2020). Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica. *Rev. UIS Ing*. 9(14), 1-18. https://doi.org/10.18273/revuin.v19n4-2020001

Scotto, V. (2025). Límites y desafíos de los grandes modelos de lenguaje para la escritura académica y científica: una revisión crítica en función del "uso experto". *Signo y Seña.* 47, 61-84

Shafique, M.N., Adeel, U., & Rashid, A. (2024). The Synergy Between Industry 5.0 and Circular Economy for Sustainable Performance in the Chinese

Manufacturing Industry. *Sustainability*, *16*(22), 9952. https://doi.org/10.3390/su16229952

Shentu, X. (2024). A review on legal issues of medical robots. *Medicine*, *103*(21), e38330. https://doi.org/10.1097/MD.0000000000038330

Silva, L.A. da, Vasconcelos, E.S., Paiva, L.F.R. de, Collaço, M.H. do V.R., Joaquim, W.M., Lima, A.D. de, Goulart, C.S., & Teixeira, E.P. (2025). Consumo energético dos data centers e a métrica da energia primária: análise global 2000–2024 e projeções até 2030. *Observatório de la Economía Latinoamericana*, *23*(10), e11812. https://doi.org/10.55905/oelv23n10-063

Smuha, N.A. (2025). *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence.* Cambridge: Cambridge University Press

Sprockel, J., Fandiño, A., Chaves, W.G., Benavides, C.O., & Diaztagle, J.J. (2023). Ensamble de redes neuronales artificiales ponderado mediante características operativas para el pronóstico de la insuficiencia cardiaca aguda. *Revista Colombiana de Cardiología*, *30*(5), 235-242. https://doi.org/10.24875/rccar.22000065

Taherdoost, H. (2024). Más allá de lo supervisado: El auge del aprendizaje autosupervisado en sistemas autónomos. *Información* , *15* (8), 491. https://doi.org/10.3390/info15080491

Tejada Díaz, N.L., Gisbert Soler, V., & Pérez Molina, A.I. (2017). Metodología de estudio de tiempo y movimiento; introducción al GSD. *3C Empresa, investigación y pensamiento crítico.* Edición Especial, 39-49.

http://dx.doi.org/10.17993/3cemp.2017.especial.39-49

Villalobos-Murillo, J., Garita González, G., & Alfaro Ramírez, B.J. (2025). Inteligencia Artificial: marco de competencias para orientar prácticas supervisadas y aplicaciones de redes neuronales en estudiantes de Computación. *Spirat. Revista Académica De Docencia Y Gestión Universitaria*, *3*(NE1), e5378. https://doi.org/10.20453/spirat.v3iNE1.5378

Voeneky, S., Kellmeyer, P., Mueller, O., & Burgard, W. (2022). *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*. Cambridge: Cambridge University Press

Walker, P.B., Haase, J.J., Mehalick, M.L., Steele, C.T., Russell, D.W. & Davidson, I.N. (2025). Aprovechamiento de la metacognición para una IA segura y responsable. *Technologies*, *13* (3), 107. https://doi.org/10.3390/technologies13030107

Wang, Y., Zhou, W., Rao, Y., & Guan, H. (2025). A Knowledge and Semantic Fusion Method for Automatic Geometry Problem Understanding. *Applied Sciences*, *15*(7), 3857. https://doi.org/10.3390/app15073857

Watson, E., Viana, T., Zhang, S., Sturgeon, B. y Petersson, L. (2024). Hacia un marco integral de ajuste personal para la alineación de valores de la IA. *Electronics* , *13* (20), 4044. https://doi.org/10.3390/electronics13204044

De esta edición de *"Inteligencia artificial: un enfoque moderno y aprendizaje profundo"*, se terminó de editar en la ciudad de Colonia del Sacramento en la República Oriental del Uruguay el 23 de febrero de 2026

Por Roberto Segundo Tejada Rodriguez, Santiago Gonzales Mesia, José Luis Castro Ullilen, Juan José Palomino Ochoa, Rosario Leonor Palomino Ochoa, María Micaela Castillo De Lima, Claudia Patricia Yon Delgado

Inteligencia artificial: un enfoque moderno y aprendizaje profundo

ISBN: 978-9915-698-71-7

www.editorialmarcaribe.es

EST. 2021 EMC
EDITORIAL MAR CARIBE