La inteligencia artificial en la educación: límites cognitivos, el rol esencial del docente y consideraciones éticas

Artificial intelligence in education: Cognitive limits, ethical risks, the essential teacher role and ethical considerations

Gonzalo Hernán Cáceres Herrera
gonza20042001@gmail.com/ https://orcid.org/0009-0001-3024-0422
Universidad Cooperativa de Colombia

Fecha de recepción: 25 de marzo de 2026
Fecha de aceptación: 14 de abril 2026
Fecha de publicación:1de julio de 2026

Favor de citar este artículo de la siguiente forma:
Cáceres Herrera, G.H. (2026). La inteligencia artificial en la educación: límites cognitivos, el rol esencial del docente y consideraciones éticas. *AULA Revista de Humanidades y Ciencias Sociales,*72 (2), https://doi.org/10.33413/aulahcs.2026.72i2.483

RESUMEN

Los modelos generativos de inteligencia artificial (IA), particularmente los grandes modelos de lenguaje (LLMs), han transformado la educación mediante la generación automatizada de contenidos, tutorías personalizadas y evaluación asistida (Holmes et al., 2019). Sin embargo, estos sistemas funcionan mediante correlaciones estadísticas sin comprensión semántica, razonamiento causal ni anclaje experiencial, lo que genera riesgos como desinformación sistemática, dependencia cognitiva y erosión del pensamiento crítico (Bender et al., 2021; Marcus, 2020). Este artículo reflexivo examina la evolución desde la IA simbólica hacia paradigmas conexionistas (Vaswani et al., 2017), analiza límites técnicos en contextos pedagógicos y cuestiona mitos impulsados por el entusiasmo tecnológico. Se evalúan impactos específicos en el aprendizaje humano —pérdida de autonomía, amplificación de sesgos, atrofia metacognitiva— y se enfatiza el rol irremplazable del docente como mediador crítico. A partir de la experiencia práctica del autor y el análisis de literatura especializada (Floridi & Cowls, 2019; Holmes et al., 2019), se establecen principios éticos para una integración responsable de la inteligencia artificial, en consonancia con marcos regulatorios emergentes como el Reglamento de IA de la Unión Europea (European Union, 2024). Se concluye que la IA potencia la educación sólo cuando queda subordinada a supervisión humana pedagógicamente informada, preservando la dimensión reflexiva del aprendizaje.

Palabras clave: *Inteligencia artificial generativa, modelos de lenguaje, pensamiento crítico, educación superior, rol docente, ética pedagógica.*

ABSTRACT

Generative artificial intelligence (AI) models, particularly large language models (LLMs), have transformed education through automated content generation, personalized tutoring, and assisted assessment (Holmes et al., 2019). However, these systems operate via statistical correlations lacking semantic understanding, causal reasoning, or experiential grounding, posing risks such as systematic disinformation, cognitive dependency, and critical thinking erosion (Bender et al., 2021; Marcus, 2020). This reflective article examines the evolution from symbolic artificial intelligence to connectionist paradigms (Vaswani et al., 2017), analyzes technical limitations in educational contexts, and challenges myths driven by technological enthusiasm. It evaluates impacts on human learning, including loss of autonomy, bias amplification, and metacognitive decline, while emphasizing the teacher's essential role as a critical mediator. Based on the author's practical experience and specialized literature (Floridi & Cowls, 2019; Holmes et al., 2019), it proposes ethical principles for responsible integration aligned with regulatory frameworks such as the European Union Artificial Intelligence Act (European Union, 2024). It concludes that artificial intelligence enhances education only when subordinated to pedagogically informed human supervision, preserving the reflective dimension of learning.

Keywords: *Generative artificial intelligence, language models, critical thinking, higher education, teacher role, pedagogical ethics.*

Introducción

La inteligencia artificial generativa redefine los procesos educativos mediante capacidades como la síntesis instantánea de información, generación de ejercicios personalizados y retroalimentación automatizada (Holmes et al., 2019). Herramientas como ChatGPT y sus sucesores han generado expectativas de transformación radical, proyectando sobre sistemas esencialmente estadísticos capacidades de comprensión humana (Bender et al., 2021; Marcus, 2020). Sin embargo, esta ilusión tecnológica oculta limitaciones estructurales: ausencia de razonamiento causal, fragilidad ante el sentido común y dependencia absoluta de patrones entrenados (Bender et al., 2021; Marcus, 2020).

El presente artículo sostiene que los LLMs constituyen herramientas complementarias valiosas pero subordinadas al juicio pedagógico humano, cuya integración indiscriminada amenaza el desarrollo de competencias metacognitivas esenciales (Kahneman, 2011). Se estructura en análisis histórico de la IA, límites técnicos específicos, riesgos educativos documentados, impacto psicológico en el aprendizaje, transformación del rol docente y marcos regulatorios aplicables educación (Floridi & Cowls, 2019; European Union, 2024; Holmes et al., 2019). La pregunta orientadora resulta crucial: ¿cómo integrar sistemas predictivos estadísticos en entornos educativos sin comprometer la formación de capacidades analíticas autónomas?

Fundamentos históricos: De la lógica formal a la estadística distribuida

La primera generación de inteligencia artificial paradigma simbólico (1956-1987) formalizaba conocimiento experto mediante reglas lógicas explícitas. Sistemas como ELIZA o MYCIN demostraron eficacia en dominios delimitados, pero revelaron limitaciones ante la complejidad combinatoria del mundo real. Los "inviernos de la IA" marcaron transición hacia paradigmas conexionistas basados en redes neuronales multicapa (Marcus, 2020).

La arquitectura transformer revolucionó el procesamiento del lenguaje mediante mecanismos de atención auto-supervisada, escalando correlaciones lingüísticas hasta capacidades emergentes en modelos masivos (Vaswani et al., 2017). Esta evolución ontológica de símbolos discretos a embeddings distribuidos (Representaciones vectoriales continuas) explica simultáneamente éxitos espectaculares en pruebas estandarizadas y fallos sistemáticos en razonamiento contrafactual o sentido común básico (Bender et al., 2021; Marcus, 2020). Desde la perspectiva educativa, esta distinción resulta fundamental para calibrar expectativas realistas sobre sus capacidades pedagógicas.

Tabla 1: Comparación ontológica de paradigmas IA

Paradigma	Representación del conocimiento	Mecanismo de aprendizaje	Explicabilidad	Escalabilidad	Aplicación educativa óptima
Simbólico	Símbolos lógicos estructurados	Codificación manual	Alta	Baja	Tutores dominio específico
Conexionista	Embeddings distribuidos	Entrenamiento masivo	Baja	Alta	Generación de contenido

Creación del autor

Límites técnicos intrínsecos de los LLMs

Los grandes modelos de lenguaje predicen secuencias mediante optimización probabilística, sin modelado semántico explícito. Tienen limitaciones estructurales que resultan críticas en contextos educativos (Bender et al., 2021; Marcus, 2020).

Por un lado, como no tienen cuerpo ni sensores que los conecten al mundo físico, les cuesta mucho incluso relaciones básicas de espacio y tiempo. Además, su aparente "sentido común" se rompe con facilidad cuando aparecen situaciones raras o poco frecuentes. También tienden a inventar datos: generan respuestas que suenan muy convincentes pero que, en realidad, pueden ser falsas o no estar respaldadas por hechos (Bender et al., 2021; Marcus, 2020). A esto se suma que se les da bien detectar patrones y correlaciones, pero son bastante malos explicando causas y mecanismos, es decir, por qué pasan las cosas. Por último, funcionan como cajas negras: no sabemos bien cómo llegan a cada decisión, lo que dificulta revisar y justificar sus juicios, algo especialmente problemático si los queremos usar para evaluar o calificar estudiantes (Floridi & Cowls, 2019; European Union, 2024).

Riesgos específicos en entornos pedagógicos

Desinformación generativa a escala

La capacidad para crear textos falsos o engañosos masivamente, que parecen completamente reales, genera un problema serio en la educación porque los estudiantes pueden tomar como ciertas informaciones inventadas por la IA, alterando cómo entienden y validan el conocimiento (su epistemología). La literatura muestra que cuando la IA produce afirmaciones falsas, pero las presenta con un lenguaje fluido y convincente, las personas —incluidos estudiantes— las aceptan fácilmente como verdaderas, sin cuestionarlas (Bender et al., 2021). En resumen, la IA no solo miente bien, sino que lo hace a gran escala, confundiendo la verdad académica y erosionando la capacidad crítica de aprendizaje.

Además, la proliferación de sistemas generativos ha facilitado la creación de contenidos sintéticos altamente convincentes, incluyendo imágenes, audios y videos falsos (deepfakes). Esta capacidad amplifica los riesgos de manipulación informativa, desinformación política y distorsión de la realidad social, afectando directamente la formación del criterio en contextos educativos (Floridi & Cowls, 2019; Holmes et al., 2019).

Dependencia cognitiva documentada

Los estudiantes que delegan sistemáticamente tareas cognitivas a los modelos de lenguaje (LLMs) experimentan un deterioro en funciones ejecutivas clave: la memoria de trabajo activa, la planificación autónoma y la metacognición evaluativa. Esto es similar al efecto del "GPS cognitivo": usar GPS para navegar constantemente reduce la capacidad mental de orientarse por uno mismo en el espacio (Sweller, 2011).

Amplificación sesgada histórica

Los datos con los que se entrenan los modelos de IA contienen prejuicios estructurales heredados de la historia: sesgo de género en recomendaciones profesionales, sesgo racial en análisis faciales, y sesgo ideológico en resúmenes de temas controvertidos (Bender et al., 2021; Floridi & Cowls, 2019). Al aprender de estos datos sesgados, la IA no solo reproduce estos prejuicios, sino que los amplifica a gran escala.

Impacto psicológico en el aprendiz

Atrofia metacognitiva progresiva

El aprendizaje efectivo, según Daniel Kahneman en su libro Pensar rápido, pensar despacio (2011), depende del Sistema 2: un modo de pensamiento lento, analítico y deliberado que requiere esfuerzo consciente para razonar, argumentar y resolver problemas complejos. En contraste, la IA proporciona respuestas instantáneas basadas en el Sistema 1 (rápido, intuitivo y automático), lo que elimina la necesidad de ese esfuerzo. Como resultado, los estudiantes practican menos la argumentación profunda y se debilita su metacognición: la habilidad esencial de reflexionar sobre sus propios procesos de pensamiento, evaluar sus ideas y detectar errores (Kahneman, 2011; Sweller, 2011).

Distorsión perceptual de realidad

Los algoritmos de recomendación de plataformas digitales (como redes sociales o motores de búsqueda) filtran el contenido que vemos según nuestros clics y preferencias previas, creando burbujas epistémicas o "filter bubbles": entornos donde solo recibimos información que refuerza nuestras creencias existentes, aislándonos de perspectivas opuestas (Floridi & Cowls, 2019). Esto genera polarización extrema, donde las personas perciben realidades completamente distintas sobre los mismos hechos (por ejemplo, un evento político visto como "victoria" por unos y "derrota" por otros). A largo plazo, distorsiona nuestra comprensión objetiva del mundo, haciendo más difícil el diálogo racional y la búsqueda de verdad compartida (Holmes et al., 2019).

Consideraciones éticas de la inteligencia artificial en educación

El uso de inteligencia artificial en contextos educativos plantea desafíos éticos fundamentales que trascienden lo técnico. La delegación de procesos cognitivos en sistemas opacos cuestiona principios como la autonomía del estudiante, la equidad en el acceso al conocimiento y la responsabilidad sobre las decisiones automatizadas (Floridi & Cowls, 2019). Asimismo, el uso acrítico de estos sistemas puede reforzar desigualdades existentes, amplificar sesgos y debilitar la formación del juicio moral. Por ello, la integración de la IA en educación requiere marcos éticos claros, transparencia en su funcionamiento y una supervisión humana constante que garantice que estas tecnologías se utilicen como herramientas de apoyo y no como sustitutos del pensamiento crítico y la deliberación humana (Floridi & Cowls, 2019; European Union, 2024; American Psychological Association, 2020)

Transformación del rol docente

El docente del siglo XXI debe actuar como orquestador de un ecosistema híbrido humano-máquina, integrando inteligentemente la

IA con el aprendizaje humano para maximizar sus beneficios y minimizar riesgos (Holmes et al., 2019). Esto implica adoptar cuatro estrategias clave:

1. Ingeniería crítica de prompts: Enseñar a los estudiantes a formular preguntas precisas a la IA y analizar sus respuestas críticamente. Por ejemplo: "Identifica tres supuestos no enunciados en esta síntesis generada por IA", entrenando su capacidad para detectar lagunas lógicas.

2. Desconstrucción comparativa: Comparar las respuestas de múltiples modelos de IA (ChatGPT, Claude, Gemini, etc.) para revelar inconsistencias, sesgos y limitaciones, fomentando el pensamiento crítico sobre la fiabilidad de estas herramientas.

3. Experiencias encarnadas obligatorias: Diseñar proyectos prácticos del mundo físico (construir prototipos, experimentos científicos, trabajo de campo) que requieran comprensión causal real, contrarrestando la desconexión de la IA puramente digital.

4. Ética aplicada a la IA: Guiar análisis sistemáticos de sesgos en los datos de entrenamiento de los modelos, discutiendo cómo estos prejuicios históricos se manifiestan en las respuestas y cómo mitigarlos en contextos educativos y profesionales.

Marco regulatorio: AI Act europeo

Reglamento (UE) 2024/1689 clasifica IA educativa (European Union, 2024).:

· Prohibida: inferencia emocional real-time menores.

· Alto riesgo: evaluación automatizada (transparencia obligatoria).

· Limitada: chatbots generales (etiquetado).

Evolución estadística sin inteligencia general

Los modelos de lenguaje grandes (LLMs) representan una evolución cuantitativa —más datos, más parámetros, más potencia computacional— pero sin salto cualitativo hacia la inteligencia artificial general (AGI), que requeriría comprensión profunda del mundo, conciencia y razonamiento autónomo verdadero (Marcus, 2020).

Su potencia estadística acelera increíblemente el acceso a información, pero amenaza las cualidades esenciales del aprendizaje humano:

· Reflexión deliberada: La capacidad de pensar lento, cuestionar premisas y construir argumentos paso a paso (Kahneman, 2011).

· Juicio ético contextualizado: Evaluar moralmente situaciones específicas considerando matices culturales, emocionales y temporales (Floridi & Cowls, 2019).

· Síntesis narrativa personalizada: Crear historias coherentes que integren experiencias personales, valores y contexto vital único (Holmes et al., 2019).

En resumen, los LLMs son aceleradores de información, no sustitutos de la inteligencia humana profunda que define el aprendizaje auténtico.

Rol transformador del docente

El docente del siglo XXI emerge como arquitecto cognitivo híbrido, diseñando entornos donde humanos y máquinas colaboren de forma sinérgica. Su misión es una integración responsable de la IA, anclada en regulaciones como el AI Act de la Unión Europea (que clasifica riesgos y exige transparencia) y guiada por pedagogía crítica que cuestiona supuestos y fomenta autonomía intelectual (European Union, 2024; Floridi & Cowls, 2019).

Esta integración convierte la amenaza de la IA en una palanca transformadora para el aprendizaje. El desafío central es educar humanos capaces de pensar con las máquinas, sin confundirlas con ellas mismas, preservando la fractura ontológica esencial: la diferencia radical entre la predicción algorítmica (basada en patrones estadísticos) y la comprensión humana auténtica (consciente, intencional y experiencial) (Kahneman, 2011; Marcus, 2020).

Conclusión

La inteligencia artificial generativa representa una evolución tecnológica significativa en el ámbito educativo; sin embargo, no constituye un sustituto de la inteligencia humana ni del juicio pedagógico. Como se ha argumentado, sus limitaciones estructurales —ausencia de comprensión, fragilidad del sentido común y funcionamiento basado en correlaciones estadísticas— implican riesgos relevantes en términos de desinformación, dependencia cognitiva y debilitamiento del pensamiento crítico.

En este contexto, el papel del docente no sólo se mantiene, sino que se fortalece como mediador crítico, garante de la reflexión y orientador del uso ético de estas tecnologías. La integración de la inteligencia artificial en educación debe realizarse bajo principios de supervisión humana, transparencia y desarrollo de competencias metacognitivas, evitando su adopción acrítica o sustitutiva.

La IA no reemplaza la inteligencia humana. Es una herramienta poderosa que debe ser utilizada con pensamiento crítico. El reto no es que la IA piense, sino que los humanos dejen de pensar.

Referencias

American Psychological Association. (2020). *Publication manual of the American Psychological Association (7th ed.)*. https://doi.org/10.1037/0000165-000

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union.*

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review,* 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning.* Center for Curriculum Redesign.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Marcus, G. (2020). *The next decade in AI: Four steps towards robust artificial intelligence.* arXiv. https://arxiv.org/abs/2002.06177

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gómez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems,* 30, 5998–6008.

Sweller, J. (2011). Cognitive load theory. *Psychology of Learning and Motivation,* 55, 37–76. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Gonzalo Hernán Cáceres Herrera

Ingeniero de Sistemas, egresado de la Universidad Cooperativa de Colombia, con formación en seguridad informática y énfasis en el análisis de vulnerabilidades y el fortalecimiento de infraestructuras tecnológicas. Cuenta con más de 15 años de experiencia en la gestión de sistemas de información, logística y optimización de procesos organizacionales.

En los últimos años ha orientado su desarrollo profesional hacia el campo de la inteligencia artificial. Sus intereses de investigación se centran en la aplicación de la IA para la optimización de procesos logísticos, la analítica avanzada y el desarrollo de sistemas inteligentes escalables.

Domina lenguajes y herramientas como Python, R, SQL y Power BI, con experiencia en el procesamiento de grandes volúmenes de datos, la ingeniería de características, la validación de modelos y el despliegue de soluciones analíticas.