

RIESGOS DE LA INTELIGENCIA ARTIFICIAL EN LA EDUCACIÓN SUPERIOR: PREMISAS PARA RESISTIR EL DETERIORO HUMANO Y COGNITIVO

RISKS OF ARTIFICIAL INTELLIGENCE IN HIGHER EDUCATION: THEORETICAL PREMISES FOR RESISTING HUMAN AND COGNITIVE EROSION

Liliana Valladares Riveroll 

Universidad Nacional Autónoma de México, México

lilianavalladares@filos.unam.mx

RESUMEN

El presente texto analiza teóricamente los resultados reportados en la literatura especializada en identificar los riesgos de utilizar la IA en la educación superior. A partir de una investigación documental, se exponen algunos de los principales usos de la IA en este nivel educativo, así como sus impactos e implicaciones en el aprendizaje, destacando algunos de sus efectos negativos sobre todo en la agencia epistémica. Derivado de la reflexión teórica, se propone un conjunto de seis premisas que, a modo de sugerencias práctico-programáticas, plantean distinciones conceptuales relevantes para el diseño de estrategias y políticas educativas enfocadas en promover una integración responsable de la IA, en la que se prevenga y evite el deterioro de las capacidades cognitivas humanas.

Palabras clave: inteligencia artificial, educación superior, riesgos de la IA, aprendizaje, inteligencia híbrida

ABSTRACT

This theoretical manuscript analyzes the reported results about the risks of using AI in higher education. Based on documentary research, the article presents some of the main uses of AI at this educational level, as well as its impacts and implications for learning, highlighting some of its negative effects, particularly on epistemic agency. From a critical reflection, a set of six theoretical premises is proposed which, in the form of practical and programmatic suggestions, raise relevant conceptual distinctions for the design of educational strategies and policies focused on promoting the responsible integration of AI, preventing and avoiding the deterioration of human cognitive abilities.

Keywords: artificial intelligence, higher education, AI risks, learning, hybrid intelligence

Introducción

Actualmente se estima que al menos 700 millones de personas utilizan los principales sistemas de Inteligencia Artificial (IA, en adelante) semanalmente (Bengio et al., 2026). En algunos paí-

ses, más del 50 % de la población utiliza la IA en salud, investigación científica, educación y otros sectores, aunque este uso es muy desigual en todo el mundo (Bengio et al., 2026).

De acuerdo con Mollema (2025), la

IA puede definirse como sistemas algorítmicos que realizan tareas semejantes a las humanas, a partir de una relación de entrada-salida (input-output). Reimers et al. (2026) entienden también a la IA como imitación de la inteligencia humana por sistemas informáticos, que ejecutan tareas como comunicarse, memorizar, razonar, aprender, reconocer patrones e imágenes, y tomar decisiones.

La imitación o simulación se refiere a que este comportamiento no es realmente inteligente ni intencional (Mollema, 2025), sino semejante a la inteligencia humana (Nyholm, 2026). Aunque, de acuerdo con Nyholm (2026), se puede asumir que poseen una forma de inteligencia y agencia artificial, Floridi (2025) ha discutido que estos sistemas no poseen inteligencia, ni intencionalidad ni autodeterminación.

Al operar mediante procesamiento computacional y de datos, exhibir diversos niveles de interactividad, autonomía y adaptabilidad, así como una orientación a objetivos, la IA puede realizar tareas e influir en su entorno en grados que van desde la agencia artefactual hasta la artificial (Mollema, 2025). Como se basa en modelos probabilísticos, la IA no genera resultados consistentes en el tiempo y puede “alucinar”, es decir, producir una respuesta plausible, pero errónea, generar afirmaciones falsas o fabricar detalles de un resultado. Al repetir una misma tarea produce respuestas o resultados diferentes, por las actualizaciones periódicas del sistema y su naturaleza probabilística (OECD, 2026). Como se entrena con un conjunto de bases de datos disponibles, sus producciones reproducen las visiones y perspectivas predominantes, sesgando culturalmente sus respuestas (Avraamidou, 2024). Por ejemplo, pueden responder correctamente 79 % de cuestiones sobre cultura estadounidense, pero solo 12 % de cuestiones sobre

cultura etíope; su rendimiento tiende a disminuir en idiomas distintos del inglés (Bengio et al., 2026). Por tanto, sus productos requieren supervisión y escrutinio humano (OECD, 2026).

La IA tiene dificultades con la planificación y progreso a largo plazo, y no puede gestionar de forma fiable obstáculos inesperados; por tanto, puede descarrilarse por errores simples durante proyectos multietapas. Su interacción con el mundo físico es un reto; por ejemplo, aún no puede integrarse con componentes robóticos para realizar tareas físicas básicas, como las domésticas (Bengio et al., 2026). Sin embargo, su uso crece en el contexto educativo.

El presente texto se cuestiona ¿Cómo se usa la IA en la educación superior? ¿Cuáles son sus riesgos, efectos e implicaciones en el aprendizaje? ¿Qué precauciones teóricas debemos mantener a la hora de diseñar políticas educativas que regulen su uso?

A partir de un análisis documental enfocado en la literatura especializada sobre los riesgos de la IA en la educación superior, el presente artículo propone un conjunto de seis premisas programáticas que plantean distinciones conceptuales que son clave para prevenir y evitar que el uso educativo de la IA sea acrítico y devenga en un deterioro de las capacidades cognitivas humanas y de su florecimiento. En un primer apartado, se exponen algunos de los principales usos de la IA en la educación superior. Posteriormente, se presentan algunas categorizaciones de los riesgos más importantes de la IA en el contexto educativo. Se destacan riesgos sistémicos para la autonomía humana; es decir, afectaciones empíricamente soportadas, en la capacidad de las personas para tomar decisiones informadas, razonar críticamente y, en suma, para el aprendizaje profundo. Estos riesgos son

fomentados por una confianza excesiva en estos sistemas y una creciente dependencia y automatización. Para resistir al deterioro cognitivo ya documentado en la literatura, finalmente se ofrece una reflexión teórica organizada en seis premisas teóricas o distinciones conceptuales que potencialmente pueden orientar el diseño de políticas educativas que prevengan que la educación superior, y su carácter crítico y humanístico, naufrague en un mundo de adopción incremental y pasiva de la IA.

Usos principales de la IA en educación superior

De acuerdo con la OECD (2026), muchos estudiantes que utilizan IA con fines académicos frecuentemente lo hacen para ser más eficientes, más que para el aprendizaje profundo. La IA se utiliza como apoyo cognitivo para producir y generar ideas (borradores, soluciones) o resolver tareas concretas —recuperar información (56 %), explicar conceptos (45 %), resumir (31 %)—; pocos estudiantes (menos del 20 %) la usan para estructurar planes de aprendizaje o monitorear su progreso. Específicamente en educación superior se reportan usos relacionados con búsqueda de información, edición, resúmenes y parafraseo, con cerca del 33 % de estudiantes usándola como compañera de aprendizaje (OECD, 2026).

En la Encuesta sobre la IA en la Educación Superior en América Latina 2026 (Digital Education Council, 2026), aplicada en 29 instituciones latinoamericanas, se reporta un uso estudiantil para el aprendizaje 6 % más alto, comparado con los resultados globales (92 % vs 86 % global), siendo el teléfono celular inteligente el dispositivo mayormente usado para el acceso digital. Predomina ChatGPT como la herramienta más utilizada (88 %), seguido de Gemini (50 %), Meta AI (29 %) y Microsoft Copilot (20 %). Entre los usos

estudiantiles predominantes, están la búsqueda de información (65%), la obtención de ideas o borradores (61 %), el apoyo en la redacción (47 %), y como compañero de aprendizaje (31 %). Es interesante que 65 % de los estudiantes expresa preocupación porque la IA vuelva el aprendizaje demasiado superficial y desaliente el pensamiento crítico y la creatividad, pues este riesgo, como se verá más adelante, se confirma cada vez más por diferentes investigaciones.

Esta misma encuesta (Digital Education Council, 2026) expone que el uso de la IA por parte del profesorado (79 %) ha aumentado 18 %, en comparación con 2025. Entre los principales usos docentes se incluyen la creación de materiales de enseñanza (76 %), producción de material multimedia (47 %), apoyo en tareas administrativas (44 %), impulso a la participación estudiantil (36 %), entre otros. Cerca del 80 % del profesorado en la región manifiesta preocupación porque los estudiantes se vuelvan excesivamente dependientes de la IA y dejen de evaluar críticamente sus productos.

La IA puede ayudar o incluso gestionar de forma independiente procesos muy diversos, desde la personalización de la enseñanza, incluyendo nuevas formas de evaluación y retroalimentación, hasta la automatización del trabajo administrativo de las escuelas (Reimers et al., 2026; OECD, 2026). Los docentes pueden predecir el rendimiento académico de los estudiantes, recomendarles recursos didácticos adecuados, proporcionar información sobre hábitos, habilidades y necesidades de estudio, evaluar automáticamente resultados de aprendizaje, proporcionando retroalimentación y sugerencias oportunas. También la IA ofrece a los estudiantes experiencias de aprendizaje más cómodas y personalizadas, incluyendo sistemas de tutoría inteligente, chatbots y múltiples herramientas

digitales (Qin y Zhang, 2025).

Según Reimers et al. (2026), la mayor parte de la IA que se utiliza actualmente en los sistemas educativos de todo el mundo es estrecha, diseñada para realizar tareas específicas, como recomendar tareas, clasificar respuestas del estudiantado o analizar patrones de asistencia. En contraste, la IA general se ve más como posibilidad teórica que realidad, pues requiere de niveles de razonamiento, abstracción y autonomía similares al de la cognición humana.

Los impactos de la IA en el mejoramiento del aprendizaje normalmente se atribuyen a ciertas capacidades específicas para la educación; por ejemplo, para la retroalimentación, la traducción de textos o el andamiaje didáctico (Dahri et al., 2025; Noroozi et al., 2025). En cuanto a la motivación, las relaciones sociales y el aprendizaje socioemocional, la OECD (2026) reconoce que su responsabilidad continúa recayendo principalmente en las interacciones sociales y diálogos humanos.

En cuanto a la retroalimentación, se ha encontrado que los estudiantes que usaron ChatGPT mejoraron un 20 % su capacidad de comprender tópicos relacionados con áreas STEM (ciencia, tecnología, ingeniería y matemáticas) (Dahri et al., 2025). Similarmente, ChatGPT puede traducir materiales complejos a las lenguas maternas del estudiantado, ampliando su accesibilidad y superando barreras lingüísticas, o bien funcionar como tutor virtual que guía el avance, clarifica dudas y elabora explicaciones personalizadas (Dahri et al., 2025). Algunos métodos de tutoría usan el cuestionamiento socrático para desarrollar conocimientos, pensamiento crítico y reflexión sobre alguna temática (OECD, 2026). A diferencia de los tutores humanos, la IA está disponible 24/7, sin restricciones geográficas u horarias. Por

ejemplo, ChatGPT ofrece andamiaje al estudiantado, reduciendo la carga cognitiva y sugiriendo recursos personalizados (Dahri et al., 2025). Yanagiura et al. (2023) exploraron la adaptación de IA a diferentes ritmos, preferencias individuales y estilos de aprendizajes, observando mejoras en el desempeño.

Otra modalidad de andamiaje que ofrece la IA son las simulaciones de escenarios profesionales reales, como el cierre de negociaciones, consultas de clientes ficticios, desarrollo de argumentos persuasivos para debates, redacción de correos electrónicos con fines profesionales. Estas simulaciones permiten que el alumnado desarrolle confianza y practique habilidades específicas de sus futuros roles (Dahri et al., 2025).

En cuanto al aprendizaje colaborativo, la OECD (2026) identifica mejoras de pequeña a mediana escala en cuatro roles principales: actuando la IA como centro o repositorio de información, generando materiales personalizados de apoyo grupal, retroalimentando el proceso y contribuyendo con los pares y/o como facilitador en tareas grupales.

A nivel de la administración y la gestión escolar, la OECD (2026) identifica usos en la generación de modelos para establecer equivalencias entre cursos y programas, agilizando tareas de admisión, orientación profesional y análisis curricular. También se le usa en evaluaciones estandarizadas de alto impacto o aplicación de exámenes a gran escala.

Fesler et al. (2026) advierten un vacío importante de investigación causal sobre IA en entornos de educación básica; gran parte de la investigación se centra en mayores de 18 años (64 % en educación superior), en condiciones restringidas (experimentos aislados, de corta duración), y con resultados a corto plazo. El 70 % de los estudios se enfoca en los impactos en estudiantes, 40 % en

docentes y 23 % en ambas poblaciones; 35 % de los estudios revisan los impactos en habilidades matemáticas, 25 % en áreas de ciencia y programación, 20 % en alfabetización y 15 % en habilidades socioemocionales. Mucha investigación compara grupos con acceso y sin acceso a IA, lo que dificulta observar los diferentes diseños pedagógicos asociados al uso de IA, los niveles de guía didáctica o procesos de interacción que configuran ambientes de aprendizaje experimentales. Son pocos los estudios que evalúan los efectos de la IA en el compromiso o participación cognitiva, la metacognición, el aprendizaje autorregulado y otras competencias a largo plazo, y la evidencia se concentra especialmente en medidas de rendimiento a corto plazo. También se necesita más investigación sobre cómo integrar eficazmente la IA en la práctica docente y cómo desarrollar la alfabetización en IA en estudiantes y docentes (Fesler et al., 2026).

Fesler et al. (2026) identifican que la IA entre estudiantes ofrece beneficios inmediatos mientras se tiene acceso activo a estas tecnologías. Sin embargo, estos beneficios a corto plazo se acompañan de una transferencia incierta cuando la herramienta es retirada, pues cuando se evalúa de forma independiente sin apoyo de la IA, los efectos son mixtos. Esto significa que, aunque la IA puede aliviar la carga cognitiva del estudiantado y fomentar experiencias positivas de aprendizaje, al mismo tiempo se puede deteriorar el pensamiento profundo. También se destaca que herramientas diseñadas con pautas pedagógicas, y que fomentan el pensamiento independiente (como los chatbots para tutorías que proporcionan razonamiento paso a paso en lugar de respuestas directas), son más prometedoras que las herramientas de IA de propósito general.

En el caso de los docentes, Fesler et al. (2026) reportan una mayor eficiencia en tareas de preparación de clases. Por ejemplo, docentes con acceso a ChatGPT dedicaron un 27% menos de tiempo a crear lecciones que sus compañeros durante las primeras cinco semanas de uso, y un 31% menos de tiempo entre las semanas seis y diez. Este uso docente de IA disminuyó con el tiempo (del 39 % al 29 %), pero el ahorro persistió, lo que sugiere que el profesorado aprende rápidamente dónde la IA aporta valor y la implementa de forma más selectiva. También la IA (como CoPilot) proporciona retroalimentación y diagnósticos periódicos y automatizados a tutores humanos, siendo especialmente eficaz para aquellos con menos experiencia y cualificación (Fesler et al., 2026).

Según cómo se utilice, la IA puede perjudicar o potenciar el aprendizaje (OECD, 2026). Su impacto podría estimarse favorable en mejorar la calidad aparente del trabajo estudiantil; en particular, su desempeño en ciertas tareas, sin que esto repercuta en mejorar su aprendizaje profundo (es decir, la adquisición duradera y transferible de conocimientos).

Como se verá más adelante, cuando la IA se usa como atajo, para evadir el esfuerzo cognitivo y evitar la dedicación necesaria para el aprendizaje, estas herramientas merman las capacidades humanas. Su integración requiere enfatizar la agencia estudiantil, para que esta sea aliada y no atajo o sustituto de la cognición humana. Esto implica integrarla bajo esquemas híbridos y en modelos pedagógicos explícitos que coloquen al ser humano en el centro, desplazando el foco de la generación de productos, a los procesos implicados. Esto debido a que los usos benéficos de IA en educación también conllevan riesgos.

Principales riesgos en el contexto educativo

De acuerdo con el Informe Internacional Sobre la Seguridad de la IA 2026 (Bengio et al., 2026), el auge de la IA se acompaña de riesgos de seguridad particulares, como los derivados de un uso indebido, problemas de control y riesgos de ciberseguridad; también hay repercusiones en los derechos humanos, la justicia, la rendición de cuentas y la privacidad. El reto para enfrentar estos riesgos estriba en que las capacidades de la IA evolucionan rápidamente, en contraste con el tiempo que lleva la recopilación y evaluación de evidencia sobre sus efectos para la sociedad; esperar a disponer de evidencia concluyente para decidir sobre el uso de una IA deja a las sociedades vulnerables frente a riesgos potenciales.

Slattery et al. (2026) propusieron una taxonomía de riesgos asociados a la IA que los clasifica según sus factores causales y los dominios donde suceden. Por factores causales, los riesgos (y su proporción documentada) pueden clasificarse según (a) la entidad que los produce: humano (39 %), IA (41 %), otra (20 %); (b) la intencionalidad: intencional (34 %), no intencional (35 %), otra (31 %); y (c) el momento de su ocurrencia: predespliegue (13 %); postdespliegue (62 %), otro (25 %). En cuanto a los dominios identificados, y su proporción, se incluyen: (a) discriminación y toxicidad (15 %), (b) privacidad y seguridad (12 %), (c) desinformación (4 %), (d) actores maliciosos y mal uso (16 %), (e) interacción humano-computadora (7 %), (f) socioeconómico y ambiental (19 %), y (g) seguridad, fallas y limitaciones del sistema de IA (26 %).

Para la educación superior resultan relevantes principalmente los riesgos en el dominio de la interacción humano-IA, relativos al exceso de confianza (incluyendo la antropomorfización y consecuente dependencia emocional y material)

y la pérdida de la agencia y autonomía humanas (desempoderamiento, erosión cognitiva y pérdida de control humano). Así también los riesgos en cuanto a desinformación, incluyendo la generación de información falsa (junto con la adopción de creencias incorrectas y falsas, y la consecuente deriva en la toma de decisiones) y la contaminación del sistema informativo (creación de cámaras de eco, erosión de la realidad y amenazas al pensamiento crítico) (Slattery et al., 2026).

Bengio et al. (2026) ofrecen una clasificación más simple y actualizada. Para ellos, los riesgos de la IA se dividen en tres categorías: (a) uso malicioso, (b) mal funcionamiento y (c) riesgos sistémicos.

En la categoría de uso malicioso, se agrupan riesgos tales como contenido generado con fines delictivos y actividades ilícitas —fraude, extorsión, difamación, creación de imágenes íntimas no consensuadas—, influencia y manipulación, ciberataques y desarrollo de armas biológicas y químicas. El riesgo de mayores repercusiones en el campo educativo es el de influencia y manipulación, resultado de la eficacia de los contenidos generados por IA de cambiar las creencias de las personas, manipular conductas y propiciar dependencia (Bengio et al., 2026).

La IA puede tener efectos manipulativos mediante la adulación servil (sicofancia) y la suplantación de identidad. Cheng et al. (2025) encontraron que 11 de los sistemas más comunes de IA aprueban las acciones de los usuarios 50 % más que los humanos. Esta adulación conlleva a una mayor preferencia y atracción por usar aquellas IA que más validan las opiniones propias, sin cuestionamiento. La sicofancia representa un riesgo de erosionar el juicio y reducir la tendencia hacia un comportamiento prosocial.

En la categoría de mal funcionamiento se agrupan riesgos relacionados a fallas de los sistemas en su operación que

resultan en disrupciones o resultados incorrectos o perjudiciosos. Se incluyen la pérdida de control (donde recuperar el sistema es extremadamente costoso o imposible) y la fiabilidad, es decir, riesgos asociados a la producción de información falsa o inventada (alucinaciones), de código informático defectuoso, de asesoramiento médico o legal engañoso, y descoordinación y conflicto en sistemas multiagente.

Finalmente, la tercera categoría refiere a los riesgos sistémicos; esto es, a aquellos que surgen del impacto de la IA en la conducta humana, las practicas organizacionales o estructuras sociales. Aquí se ubican impactos en el mercado laboral, derivados de la automatización de tareas cognitivas, especialmente en el trabajo intelectual fácilmente sustituible (redacción, traducción), y que pueden reducir significativamente el empleo y los salarios. También se ubican los riesgos para la autonomía humana. Estos últimos son particularmente relevantes en educación superior, en tanto que la IA puede afectar la capacidad para tomar decisiones informadas y actuar en consecuencia (Bengio et al., 2026). Confiar en exceso en la IA puede debilitar las capacidades de razonamiento crítico y fomentar el sesgo de automatización, llevando a los usuarios a pasar por alto problemas que el sistema automatizado no detecta, o a actuar siguiendo consejos incorrectos, por ejemplo, en ámbitos como aviación o monitoreo de tareas (Bengio et al., 2026).

La IA y la autonomía epistémica

La autonomía puede entenderse como la capacidad efectiva de fijar objetivos que reflejen los propios valores y orientar hacia estos la propia conducta. Comprende las nociones de autenticidad (sostener valores, objetivos y motivos propios), y de agencia (capacidad y libertad de actuar y ejecutar las propias

decisiones) (Bengio et al., 2026). Yan, Pammer-Schindler et al. (2025) suman a la autonomía el concepto de compromiso epistémico, entendido como la disposición proactiva del estudiantado y su capacidad para cuestionar críticamente, validar y contextualizar el contenido generado por IA, fundamental para su integración educativa significativa. Por su parte, la OECD (2026) entiende como agencia, particularmente agencia docente, la capacidad activa y los incentivos del profesorado para tomar decisiones y ejercer influencia en su práctica profesional.

Cada vez hay más evidencia de que la dependencia de la IA para realizar tareas cognitivas afecta negativamente la capacidad de pensamiento crítico y la memoria (Bengio et al., 2026). Un ejemplo alarmante puede encontrarse en un estudio que informó que la capacidad de los médicos para detectar tumores (adenomas por colonoscopia) sin apoyo de IA fue aproximadamente un 6 % inferior, tras varios meses de haber sido expuestos a tareas de diagnóstico asistido por IA (Budzyń et al., 2025). Este estudio ha sido de los primeros en mostrar que la exposición a la IA podría tener un impacto negativo en el desempeño médico, particularmente en su descalificación para diagnosticar a pacientes (Budzyń et al., 2025). En otro estudio, estudiantes que confiaron en exceso en ChatGPT, mostraron habilidades de argumentación más débiles que aquellos que usaron métodos tradicionales (Yan, Greiff et al., 2025). Darvishi et al. (2024) encontraron cómo las ganancias significativas en la calidad de la retroalimentación entre estudiantes pares, usando una herramienta de IA, no se sostuvo cuando la herramienta fue retirada. El fomento de prácticas no guiadas, reducidas a solicitar a la IA respuestas o soluciones, dañan la reflexión y la metacognición (OECD,

2026). Similarmente, se ha mostrado cómo el uso frecuente de IA para la escritura argumentativa, en la que estudiantes delegan la revisión de sus propios textos usando rúbricas, puede promover “pereza cognitiva” y dependencia. Con esto se dañan las habilidades de autorregulación, incluyendo la autoevaluación, la planeación y revisión autónomas, así como el compromiso con los materiales de estudio. La disminución del sentido de control y autoeficacia repercute también en un decrecimiento de autonomía y de motivación intrínseca (Yan, Greiff et al., 2025; OECD, 2026).

Esto se relaciona con la teoría de la descarga cognitiva, según la cual, cuando una persona libera recursos cognitivos delegando tareas cognitivas en sistemas o personas externas, se mejora la eficiencia. No obstante, hay también un riesgo de reducir la propia implicación cognitiva y, por tanto, la capacidad de actuar con autonomía (Fesler et al., 2026).

Coeckelbergh (2026) también identifica la disminución de la agencia epistémica como uno de los principales riesgos de la IA. Por agencia epistémica, este autor entiende el control que un agente puede ejercer sobre sus creencias, incluyendo su revisión y formación. La agencia epistémica es condición necesaria de la responsabilidad moral, pues a partir de las creencias se orientan las acciones.

Coeckelbergh (2026) reconoce tres tipos de influencias que dificultan la agencia epistémica, en la medida en que obligan a las personas a esforzarse más para discernir entre distintas creencias: (a) manipulación directa de creencias (mediante propaganda que se aprovecha del engaño, creando ambientes epistémicos que dificultan distinguir entre verdad y falsedad); (b) burbujas epistémicas (estructuras epistémicas sociales en las cuales, a través de filtros algorítmicos y de personalización, se exclu-

yen y desacreditan voces, informaciones y argumentos relevantes y opuestos a los propios, generando un ambiente de desconfianza entre miembros de una burbuja); y (c) fallas del conocimiento estadístico (alucinaciones o sesgos de la IA de usar conocimientos más fácilmente accesibles para convertirlos en norma).

Para Coeckelbergh (2026), la IA dificulta el desarrollo de virtudes epistémicas porque crean sociotécnicamente una pereza epistémica. Esta puede crecer en la medida en que se prefiere usar estas tecnologías, en lugar de fortalecer el razonamiento necesario para actualizar las creencias propias a la luz de la nueva evidencia. La IA participa así de formas de injusticia y opresión epistémica al obstaculizar, de modo persistente e injustificado, la capacidad de utilizar recursos epistémicos compartidos de forma persuasiva, afectando negativamente la producción de conocimiento.

Küçükuncular y Ertugan (2026), desde una interpretación marxista, documentan experiencias de alienación entre 395 docentes de educación superior en Chipre, derivadas de las transformaciones del trabajo académico impulsadas por IA. Esta, apuntan los autores, no es solo un apoyo institucional neutral, sino que es un instrumento de gobernanza algorítmica que, mediante la mayor cuantificación, monitoreo y optimización, impacta la labor educativa. La alienación docente fue experimentada de modo profundo por docentes, al sentir una disminuida sensación de propiedad y autoría sobre los resultados de enseñanza (materiales curriculares y didácticos, evaluaciones, rúbricas, etc.). También se hizo patente una alienación sobre la identidad profesional, resultado de las presiones impuestas por las prácticas de gestión algorítmica, control y estandarización en los entornos de educación superior. Al disminuir las interacciones humanas directas, a través

de plataformas, aplicaciones o pizarras de monitoreo, el profesorado enfrentó tensiones alrededor de la erosión de su autonomía profesional y de sus capacidades creativas y relacionales interpersonales.

Por otra parte, Gerlich (2025) documentó afectaciones negativas de la IA específicamente en el pensamiento crítico, como capacidad de analizar, evaluar y sintetizar información para tomar decisiones razonadas. Al delegar tareas cognitivas a una herramienta externa, se reduce el compromiso cognitivo para evaluar información. A esta disminución se le conoce como pereza cognitiva. Los participantes más jóvenes mostraron una mayor dependencia de la IA y puntuaciones más bajas en pensamiento crítico en comparación con los participantes de mayor edad. Entre más frecuente el uso de IA, menores son las capacidades de pensamiento crítico y mayor es la descarga cognitiva, delegando procesos de memoria, atención, y toma de decisiones. Para contrarrestar estos costos, Gerlich (2025) advierte la necesidad de diseñar ambientes y estrategias educativas que promuevan una interacción crítica con la IA, y un compromiso cognitivo de los usuarios.

Fan et al. (2025) compararon experimentalmente la motivación, los procesos de aprendizaje autorregulado y el rendimiento académico de 117 estudiantes universitarios frente a una tarea de escritura. Un primer grupo se apoyó en ChatGPT; un segundo grupo en un experto humano; y un tercer grupo no tenía apoyo adicional. El grupo apoyado con ChatGPT obtuvo un mejor rendimiento en tareas a corto plazo; no obstante, su motivación intrínseca y la ganancia y transferencia de conocimientos no difirieron del resto. Lo que sí fue particularmente destacable es que el uso de ChatGPT fomentó dependencia y potencial “pereza metacognitiva”, la cual, a su vez, puede obstaculi-

zar la capacidad de autorregulación y de involucrarse en un aprendizaje profundo. La IA pueden optimizar el desempeño, pero a costa de un desarrollo genuino de habilidades cognitivas humanas.

Para Fan et al. (2025), la pereza metacognitiva refiere a la evitación habitual del esfuerzo cognitivo deliberado que ocurre cuando el estudiantado depende excesivamente de los resultados o de la facilitación generados por la IA. Esta dificulta que se activen procesos metacognitivos más profundos, se opte por una toma de decisiones menos esforzada y más intuitiva, socavando la autorregulación y el monitoreo cognitivo, centrales en el aprendizaje.

Fan et al. sostienen que es muy importante no confiar ciegamente en la IA y estimular un uso acompañado tanto de la motivación intrínseca del estudiantado, de un andamiaje para el aprendizaje activo, como de impulsar la participación humana en procesos metacognitivos (evaluación, monitoreo y orientación).

Shaw y Nave (2026) documentan un fenómeno que profundiza la pereza cognitiva y metacognitiva, y provoca lo que llaman la “rendición cognitiva”. Según estos autores, la cognición artificial —apoyada en herramientas externas como la IA— puede complementar o reemplazar los procesos internos, introduciendo nuevas vías cognitivas. La rendición cognitiva ocurre cuando se adoptan los resultados de la IA con mínimo escrutinio, anulando la intuición y la deliberación humanas. Los autores reportan que aquellos participantes con mayor confianza en la IA, un deseo de facilidad cognitiva y un pensamiento analítico menos laborioso (menor reflexión cognitiva e inteligencia) muestran una mayor rendición cognitiva. Esta rendición reconfigura su autonomía, compromiso y responsabilidad epistémicas, generando no una automatización, sino una disposición

más amplia de dependencia epistémica. La salida virtuosa de esta renuncia a la agencia radica, según los autores, en crear y cultivar rutas híbridas IA-humanos, en la que se promueva la colaboración y se conserven procesos cognitivos humanos como verificar y racionalizar. Asimismo, muestran que los incentivos y la retroalimentación pueden atenuar parcialmente la rendición cognitiva.

Similarmente, Yan, Pammer-Schindler et al. (2025) apuntan a que una integración efectiva de la IA requiere diseños instruccionales deliberados, cuidadosamente estructurados para fomentar el desarrollo metacognitivo y epistémico continuo del estudiantado.

Por su parte, Avraamidou (2024) considera la IA en la educación como una expresión de colonización tecnocrática y un vehículo para la exclusión epistémica y reproducción de ideologías neoliberales. Como contrapeso, esta autora desarrolla una postura crítica basada en una pedagogía de la IA feminista y relacional, que procura el cuidado, la justicia y la ética relacional, y que se centra en la corporalidad, la capacidad de acción y la resistencia. Identifica cinco ejes de perjuicios de la IA: colonización epistémica, injusticia ética y social, monocultura algorítmica, degradación ambiental y empobrecimiento teórico.

Chatzichristofis (2026) contraargumenta estas críticas. En lugar de rechazar la IA como inherentemente opresiva o verla como amenaza tecnocrática, este autor argumenta que la IA puede apoyar un aprendizaje inclusivo cuando se sitúa en contextos dialógicos, culturalmente sensibles y liderados por el profesorado. La capacidad de acción y de mediación del profesorado para configurar coagencias con la IA puede reorientar su uso hacia objetivos centrados en la justicia. Para lograrlo, se requiere una integración pedagógica relacional, crítica, colabora-

tiva, bajo control humano, pero también pluralista en sus métodos y mediada por valores humanos y andamiajes éticos.

Aunque la evidencia de los riesgos y perjuicios al proceso educativo y de aprendizaje sigue creciendo, aún es contradictoria, por lo que es necesario investigar más para comprender los efectos a largo plazo de las interacciones humanos-IA. Con lo reportado hasta el momento, es evidente que la IA pone en riesgo a la autonomía humana, y que justamente el ejercicio y cultivo -más no la sustitución tecnológica- de esta capacidad, frente a la adopción acrítica de la IA, es lo que asegurará la defensa y florecimiento de la agencia humana. Para lidiar con estos riesgos, se destaca la alfabetización en IA, de la que se hablará más adelante, así como la propuesta de salvaguardias técnicas y resiliencia social para la gestión de daños (Bengio et al., 2026; Coeckelbergh, 2026).

Alternativas precautorias: premisas teóricas para resistir

Frente al conjunto de riesgos, particularmente de aquellos asociados al deterioro y de capacidades epistémicas, a continuación se proponen un conjunto de premisas derivadas de la reflexión teórica conceptual, que, a modo de sugerencias práctico-programáticas, buscan orientar el diseño de políticas educativas enfocadas en impedir que la educación superior naufrage en un mundo de deshumanización, desarrollo y adopción creciente de IA.

1. El aprendizaje no equivale al rendimiento (performance)

Yan, Pammer-Schindler et al. (2025) sostienen que mucha de la investigación reciente sobre IA y educación se ha centrado predominantemente en analizar resultados y efectos positivos relacionados con el rendimiento académico, pero más allá de las mejoras en eficiencia y

escalabilidad, hay que responder interrogantes cruciales sobre sus implicaciones más profundas, ya sea cognitivas, epistémicas, socioemocionales y éticas en estudiantes y docentes.

Se debe distinguir entre el aprendizaje y el rendimiento. El rendimiento se considera una medida fundamental del éxito estudiantil y normalmente se refleja en resultados a corto plazo, como las calificaciones de exámenes (Dahri et al., 2025). El aprendizaje, en cambio, se centra en la retención a largo plazo y la capacidad de aplicar el conocimiento en el futuro (Kirschner et al., 2025). Implica cambios durables en el conocimiento y el comportamiento, que resultan de una experiencia y que comprenden habilidades de retención y transferencia a diferentes contextos (Yan, Greiff et al., 2025).

La integración de IA puede mejorar el rendimiento del estudiantado, por ejemplo, reduciendo la carga cognitiva, pero Yan, Greiff et al. (2025) advierten que estos usos no promueven los procesos cognitivos y metacognitivos profundos que el aprendizaje de alta calidad requiere; por el contrario, existe el riesgo de que estos se socaven. Un alto rendimiento no siempre implica aprendizaje, pues su medida es transitoria y dependiente del contexto. En muchos estudios, las mejoras en el rendimiento disminuyen cuando se retira la asistencia de la IA, apuntando a que las ganancias medidas en el rendimiento no reflejan un mayor aprendizaje (Yan, Greiff et al., 2025; Fesler et al., 2026).

Para lograr un aprendizaje más profundo y duradero con apoyo de IA, es necesario entender cómo estas tecnologías podrían equilibrar, al mismo tiempo, la carga cognitiva (el esfuerzo mental requerido durante las tareas), la metacognición (la conciencia y regulación de los propios procesos de pensamiento) y la autoeficacia (las creencias de los estu-

diantes sobre sus habilidades) y traducirse en mejoras sostenidas en el aprendizaje (Yan, Greiff et al., 2025).

Como afirma Barcaui (2025), tener presentes las bases teóricas y los principios fundamentales del aprendizaje humano es más importante que nunca en la era de la IA. Kirschner et al. (2025) han identificado un conjunto de diez ilusiones pedagógicas (instructional illusions) que subestiman aspectos clave de la enseñanza y el aprendizaje, ocultando verdades más profundas sobre lo que funciona y por qué; es decir, son suposiciones que no revelan la realidad completa y fundamentada de cómo se aprende. Las diez ilusiones pedagógicas discutidas por Kirschner et al. (2025) han sido implícitamente adoptadas y transferidas, en buena parte, a los contextos de investigación y uso educativo de IA, por lo que se necesitan hacer explícitas para desenmascarar interpretaciones erróneas.

Fesler et al. (2026) reportan cómo la IA puede ayudar al alumnado a completar tareas con mayor éxito en el momento, pero esas mejoras no siempre persisten en el modo independiente. Por ejemplo, Barcaui (2025) evaluó el impacto de ChatGPT en la retención de conocimientos a largo plazo en estudiantes universitarios. Los estudiantes que usaron ChatGPT obtuvieron puntuaciones significativamente más bajas en la prueba de retención (57.5 % de respuestas correctas) en comparación con quienes estudiaron de forma tradicional (68.5 % de respuestas correctas). El autor advierte que el uso ilimitado de ChatGPT perjudicó la retención a largo plazo, probablemente al reducir el esfuerzo cognitivo que sustenta la memoria a largo plazo. La asistencia de la IA puede facilitar el aprendizaje inicial, pero socavaría los procesos que requieren esfuerzo para un aprendizaje sólido.

Esta reducción de la retención es

acorde con la teoría de la descarga cognitiva y la propuesta de las llamadas dificultades deseables. Las dificultades deseables son condiciones de aprendizaje que el alumno suele experimentar como más exigentes, pero que tienen un efecto positivo en el aprendizaje y la transferencia de conocimientos y habilidades. Las condiciones de aprendizaje que favorecen una mejora rápida del rendimiento a menudo no favorecen la retención y la transferencia a largo plazo, mientras que las condiciones que plantean retos y ralentizan el ritmo del aprendizaje suelen optimizar la retención y la transferencia a largo plazo (Kirschner et al., 2025).

La consolidación de la memoria requiere tiempo y esfuerzo cognitivo. Como apunta Barcaui (2025), a diferencia de una calculadora o un cuaderno, que delegan tareas específicas o almacenamiento de memoria, la IA puede delegar procesos cognitivos completos, incluyendo la comprensión, la síntesis de información e, incluso, aspectos del pensamiento crítico. Usar la IA como muleta cognitiva agiliza el completar la tarea con un procesamiento de conceptos más superficial, creando una ilusión de competencia. Esta ilusión o “competencia prestada” consiste en que la IA puede proporcionar estructura, vocabulario y andamiajes de razonamiento que aumentan la falsa sensación de dominio durante el estudio, sin necesariamente actuar a nivel de las estructuras cognitivas, reforzando las huellas de memoria mediante la recuperación o la elaboración generada por el alumno (Barcaui, 2025). En contraste, en el aprendizaje tradicional se incluyen estrategias como relectura, autoevaluación, repaso espaciado y elaborativo, actividades que fortalecen la consolidación mediante la recuperación de información.

Noroozi et al. (2025) refieren cómo, aunque falta exploración empírica, a los impactos educativos de la IA no los de-

termina la herramienta tecnológica en sí misma, sino su abordaje pedagógico y su regulación ética. La adopción de la IA en la educación superior requiere fundamentarse pedagógicamente, centrarse en los humanos y tener un propósito educativo.

La investigación sobre usos y efectos educativos de IA debe incluir más estudios longitudinales y a largo plazo e informarse de las teorías del aprendizaje, la psicología educativa y ciencias cognitivas. Asimismo, debe clarificar las interacciones cognitivas humanos-IA, examinando su influencia en la codificación, consolidación y recuperación de información, distinguiendo entre las mejoras en el rendimiento y en el aprendizaje (Yan, Greiff et al., 2025). Para evitar los daños a las capacidades cognitivas, estas tecnologías no deben nunca integrarse como reemplazo o sustituto de los esfuerzos cognitivos, sino servir para aumentar o transformar el aprendizaje humano (Yan, Greiff et al., 2025; Barcaui, 2025). Esto requiere diseñar una arquitectura pedagógica que considere, entre otros, el marco de las dificultades deseables para un aprendizaje duradero (Kirschner et al., 2025).

2. La potencial trampa de la IA trastoca el proceso de educación racional y crítica

Ciocchetti (2026) propone el concepto de la “trampa de la IA” para referir a la tendencia a depender de la IA como atajo que elude el esfuerzo humano, obstaculizando la formación del carácter y la presencia encarnada y relacional necesarios para el florecimiento. Caer en esta trampa significa perder oportunidades pedagógicas únicas para el desarrollo humano, por lo que el autor sugiere desconectarse de la IA para decidir o actuar frente a situaciones complejas.

La educación no es simplemente la adquisición de conocimiento, sino un

proceso de transformación en el cual un sujeto humano se convierte en persona, en alguien —cultural y simbólicamente— cada vez distinto de lo que —biológicamente— es. Esta conversión (*becoming*) o subjetivación se remonta a la concepción de la educación como un proceso de humanización, como lo han propuesto autores clásicos como Platón, Kant o Freire (Chutorański, 2026). La activación de este potencial humano no es espontánea, sino que normalmente requiere de una estructura social y de relaciones pedagógicas que expanden oportunidades, impulsan la emancipación, construyen comunidades, cultivan valores; estas relaciones ocurren en diferentes espacios sociales compartidos y a lo largo de toda la vida, desde el kínder hasta la universidad (Chutorański, 2026).

Para Kant (2003), los seres humanos son perfectibles y la acción educativa progresa de la mano de una concepción ideal de lo que la humanidad puede llegar a ser. En la mirada kantiana, mediante la disciplina, la instrucción y la formación se puede lograr el desarrollo de las capacidades y facultades propiamente humanas. Particularmente con la formación (*bildung*), entendida como una educación moral y cívica, el ser humano se forma para vivir con rectitud y obrar en libertad, aprender a pensar, gobernar su voluntad y usar su razón para elegir fines buenos (respeto a la dignidad humana, justicia, igualdad, libertad, entre otros). El desarrollo de las capacidades racionales del individuo posibilita su formación como ser autónomo, es decir, capaz de poseer un pensamiento propio y darse a sí mismo las máximas en las que se fundan sus acciones. Por consiguiente, la mirada kantiana subraya que el carácter de la educación en tanto formación, no se reduce a un mero proceso de instrucción o capacitación, sino que busca formar sujetos con autonomía, que guar-

dan una actitud crítica ante las normas y principios que regulan la sociedad, y que actúan para transformar y perfeccionar su ser y sus instituciones (Beade, 2011).

Con la introducción de la IA en la educación, se corre el riesgo de ceder este desarrollo de la autonomía a un sistema tecnológico, renunciando con ello a la meta fundamental kantiana del proceso educativo como camino de humanización y progreso de las instituciones sociales. Esta renuncia a la autonomía, y su delegación a la IA, se traduciría en una mentalidad heterónoma pasiva, ciega, que no somete a ningún juicio de la razón las contradicciones o apariencias, dependiente y sometida a cualquier tipo de autoridad ajena y arbitraria, como la que supone un algoritmo.

Ahora bien, con raíces kantianas, pero con el sesgo ideológico heredado del marxismo, Costa y Murphy (2026) han cuestionado el papel de la IA en el desarrollo de una educación crítica en el sentido propuesto por Freire. Aunque pueda ser criticable su postura, estos autores sostienen la importancia de re-examinar cómo se concibe teóricamente la educación ante la integración de IA y cómo se organiza pedagógicamente esta incorporación tecnológica. Reconocen que, aunque su uso puede traer ganancias y una promesa de liberación de tareas intensivas, también conlleva amenazas a la educación crítica en el sentido freireano.

Para Freire, la educación como acto político y liberador promueve la capacidad de cuestionar el orden establecido, denunciar las injusticias y reconstruirlo conforme a valores como la justicia social y la solidaridad. La educación no se reduce al desarrollo de habilidades técnicas, sino que provoca una emancipación y empoderamiento intelectual en el que se cultiva la concientización y la autonomía. En concreto, señalan Costa y

Murphy (2026), los elementos de la educación crítica que Freire destacaba para ejercer la libertad responsable incluyen (a) procesos de problematización, (b) diálogo crítico y (c) actos de conocer. La integración de IA en la educación trastoca estos tres elementos al configurar una noción reducida del proceso educativo.

Costa y Murphy (2026) sostienen que la autonomía y la emancipación, que son beneficios a largo plazo de una educación crítica, se erosionan cuando no se transita por los tres procesos de problematización, diálogo crítico y praxis asociada a los actos de conocimiento. En su lugar, llegar a respuestas preestablecidas generadas por IA provoca una ilusión de eficiencia y de liberación del trabajo mental. No obstante, el aprendizaje se instrumentaliza como una acción solitaria y competitiva, que fomenta la dependencia tecnológica y la domesticación intelectual, en aras de lograr una mayor eficiencia.

3. Lo que es social no puede reducirse ni automatizarse bajo la seducción de la IA

A diferencia de las respuestas basadas en algoritmos que la IA puede aportar, Coeckelbergh (2025) sostiene que la vida social se centra en el significado y la hermenéutica; por tanto, la participación en sociedad requiere variedad y flexibilidad en las respuestas. Centrada en la interpretación y el diálogo, la dimensión hermenéutica de lo social se enfoca en cómo las personas se comprenden a sí mismas, desentrañando una red de significados más amplia, que configura un trasfondo u horizonte de vida que posibilita la comprensión e interpretación. Según Coeckelbergh (2025), para una máquina o tecnología social resulta imposible captar esta red de significados o trasfondo de significados, pues es una dimensión típicamente humana. Las interacciones naturales y encarnadas entre

humanos dependen mucho del contexto y de la persona, incluyen improvisación, empatía, entendimiento profundo, y no son solo repeticiones estereotipadas de un rol como propone la IA.

En contraste con los humanos, los sistemas artificiales carecen: de un entendimiento del contexto cultural, incluyendo la interpretación de situaciones emocionales, morales y sociales; de conocimiento tácito encarnado en sensaciones, percepciones, movimientos y experiencias sensoriomotoras para adaptarse a diferentes entornos; de razonamiento moral y empatía para establecer conexiones socioemocionales, así como juicios éticos, en función de una apreciación hermenéutica contextualizada (Lutz et al., 2025).

La generación de conocimiento emerge de interacciones humanas situadas que, como apunta Floridi (2025), conllevan formas de agencia colectiva que van más allá de las capacidades individuales que puede reproducir un algoritmo, y en las que surgen actuaciones de cooperación, cuidado y colaboración, dentro de sistemas sociales estructurados, basados en marcos culturales, lingüísticos e institucionales complejos. En estos sistemas humanos no solo se negocian evidencias y argumentos, sino que también se construyen vínculos afectivos y de confianza que van configurando una memoria humana distribuida. Esta memoria no depende de una arquitectura externa de datos e información explícita y declarativa como la IA, sino de una arquitectura social y encarnada que se sostiene en prácticas colectivas, de acompañamiento, mentoría y rutinas compartidas que preservan el carácter tácito de la experiencia humana y facilitan su transferencia intergeneracional, mediante procesos relacionales de socialización (Collins, 2025; Ozioma et al., 2026).

4. La defensa de lo humano implica no renunciar a la responsabilidad hermenéutica y narrativa ni transferirla a la IA

Para Coeckelbergh (2023), la responsabilidad hermenéutica es principalmente un tipo de responsabilidad en primera persona. Es una responsabilidad relacional y narrativa, puesto que se debe comunicar el sentido de algo a los demás, y esto generalmente se presenta en forma de narración. No es condenable usar tecnologías digitales, así como diversos medios y prácticas culturales —escritura, texto, libros— para ejercer esta responsabilidad hermenéutica a través de la construcción de narrativas que dan sentido al mundo, pero este uso de las tecnologías, sobre todo de la IA, no debe traducirse en una renuncia a una responsabilidad propiamente humana, que “es y sigue siendo nuestra responsabilidad” (Coeckelbergh, 2023, p. 2442).

La narratividad puede ser una actividad distribuida y compartida por autores y lectores humanos y no-humanos; en tal sentido, la IA puede ser cocreadora de narrativas, pero Coeckelbergh (2023) sostiene que la responsabilidad de dar significado recae en el ser humano:

La IA forma parte de nuestras narrativas y ayuda a moldearlas, pero es nuestra responsabilidad definir su papel como cocreadores y depende de nosotros el lugar y el rol que les asignamos en las narrativas que co creamos. En definitiva, es nuestra responsabilidad decidir qué narrativas queremos escribir (o coescribir), incluyendo narrativas sobre la IA, con la IA y, a veces, en contra de la IA. (p. 2246)

Es también responsabilidad humana revisar, reconfigurar y reescribir las historias que la IA construye, pues su papel de coautora no significa la aceptación acrítica de sus productos, ni que estas producciones posean un significado en un sentido profundo, humano y socioe-

xistencial, pues carecen de conciencia y de la experiencia de la auténtica narratividad. De otro modo el riesgo es “quedar a merced de las identidades y las historias que estas tecnologías (y sus diseñadores y empleadores) cocrean. Comprender y evaluar el papel narrativo de la IA es, por lo tanto, una responsabilidad tanto normativa como hermenéutica” (p. 2246). La toma de sentido requiere de un conocimiento tácito, implícito y encarnado que no siempre puede formalizarse, así como de la transformación de una comprensión inicial hacia una nueva comprensión, la cual se produce como resultado de la experiencia humana, la interpretación y la subjetividad, que son tareas hermenéuticas que no fácilmente pueden ser sustituidas por la IA. Coeckelbergh (2023) reitera que la IA puede ser considerada coautora, pero que, dadas sus limitaciones actuales (falta de conciencia, subjetividad y experiencia), estas tecnologías no deberían sustituir ni reemplazar el trabajo hermenéutico de los seres humanos, que son quienes experimentan y crean significado. Como seres-en-el-mundo, el proceso hermenéutico de construcción de significado siempre debe producirse a través de ellos, y alcanzar en ellos su culminación; y son los seres humanos quienes deben definir el papel que tendrá la IA en estas historias y no al revés. La confianza epistémica debe seguir recayendo en los humanos, por su distintiva naturaleza social, moral y hermenéutica.

Hahne y Schmoelz (2026) denuncian la tendencia creciente a antropomorfizar a la IA y tratarla como receptora u objeto de confianza, pues esto conlleva una deshumanización y una distorsión conceptual. Para estos autores, la confianza es un concepto moral intrínsecamente humano —basado en la vulnerabilidad, la autonomía y el reconocimiento recíproco— que no puede transferirse a las máquinas. La

confianza es una relación intersubjetiva producto de lo social, que implica una dimensión no cuantificable, resultado de una dialéctica entre la autoconfianza y la confianza en la comunicación humana. La capacidad de confiar en uno mismo fundamenta la capacidad de confiar en los demás, mientras que las relaciones interpersonales basadas en la confianza nutren y sostienen la autoconfianza; es decir, una refuerza a la otra. La autoconfianza se refiere a la capacidad cultivada de confiar en las propias facultades de percepción, juicio y discernimiento ético, y enfatiza la importancia de escucharse a uno mismo, defender la propia voz, confiando en los propios valores, intuiciones y experiencias internas. Ceder esta capacidad y confiar en que la IA tome decisiones a nombre de uno mismo no solo generan una creciente dependencia de los usuarios hacia sistemas opacos, sino que socavan la autonomía y el pensamiento crítico.

La confianza se desarrolla y cultiva también a través de la interacción entre la integridad personal y el reconocimiento social, pues quien confía necesita creer en la buena voluntad del depositario de su confianza, asumiendo que no tiene malas intenciones hacia él. Esta dimensión moral de lo social (y viceversa) se pierde en la IA, pues no se puede afirmar que los sistemas de IA tengan una buena voluntad, más allá de los parámetros con los que fueron programados. La IA opera basada en algoritmos y datos, y no en una conciencia humana ni en una autonomía moral, necesarias para establecer relaciones de confianza; por tanto, la IA no puede participar de un reconocimiento recíproco y es incapaz de experimentar o responder a la vulnerabilidad (Hahne y Schmoelz, 2026). Aprender a no confiar ciegamente en la IA requiere de diseños pedagógicos que favorezcan el cultivo de la confianza intersubjetiva y la autocon-

fianza en la propia intuición, voz, valores y experiencias. Mediante la reescritura, el rechazo y la resignificación activa de lo que la IA genera se puede preservar la diversidad de identidades narrativas de los usuarios, evitando la creciente homogeneización de resultados producidos por IA, los cuales exhiben textos cada vez más repetitivos, circulares o incluso degradados (Ozioma et al, 2026).

5. La IA puede reconceptualizarse como inteligencia híbrida

Además de Coeckelbergh (2023), autores como Yan et al. (2025b) o Cukurova (2025) proponen considerar la IA como colaboradora activa en el aprendizaje. Esta perspectiva reorienta los entornos colaborativos tradicionales para priorizar el diálogo, la regulación mutua y la co-construcción dinámica del conocimiento entre humanos y tecnologías.

Estos ambientes de inteligencia híbrida deberían equilibrar cuidadosamente las interacciones dirigidas por IA con la participación humana, de modo que sea posible siempre preservar y mantener la autonomía humana en la responsabilidad cognitiva compartida. Este equilibrio será adecuado si se logran aprovechar las capacidades de la IA como apoyo cognitivo, sin eclipsar la participación humana en la construcción del conocimiento, la negociación de significados y la reflexión (Yan, Pammer-Schindler et al., 2025).

Cukurova (2025) cuestiona la conceptualización restrictiva predominante de la IA como herramienta educativa, y plantea pensarla como componente de una inteligencia híbrida humano-IA. Distingue tres formas de conceptualizar la IA en educación: como una externalización de la cognición humana, como internalización de modelos de IA que influyen en modelos mentales humanos y como una extensión de la cognición humana mediante acoplamientos de inteligencia

híbrida humano-IA.

En la primera conceptualización, algunas tareas humanas se definen, modelan y reemplazan mediante la IA como herramienta. En la segunda conceptualización, la IA puede utilizarse para ayudar a los humanos a modificar sus representaciones y tareas, internalizando sus modelos. En la extensión de la cognición humana, la IA extiende la cognición humana como parte de sistemas de inteligencia híbrida entre humanos-IA. En esta variante, la inteligencia emergente es sinérgica, es decir, superior a la suma de la inteligencia de cada agente, tanto humano como artificial (Cukurova, 2025). Fan et al. (2025) definen la inteligencia híbrida como resultado de la combinación de la inteligencia humana y artificial, la cual potencia el intelecto y las capacidades humanas en lugar de reemplazarlas, permitiendo alcanzar objetivos que sin tal combinación serían inalcanzables.

Ahora bien, desarrollar sistemas de inteligencia híbrida es un proceso creciente, intencionado, informado por la evidencia y centrado en los humanos, que requiere tanto de un conjunto de competencias humanas que los operacionalicen, como de innovaciones significativas en los sistemas educativos. Cukurova (2025) reconoce la necesidad de una formación de las personas sobre IA, que les permita un uso seguro, eficaz y ético, así como el impulso de sistemas educativos innovadores enfocados en desarrollar procesos (no productos), con perspectiva humanista más que eficientista.

6. La alfabetización en IA requiere ser cultivada con un sentido ético y social

La alfabetización en IA es una competencia multidimensional que incluye el conocimiento de cómo funciona la IA y cómo desarrollarla, cómo interactuar y usarla como coagente, y la capacidad de discernir su impacto ético en las socie-

dades (Reimers et al., 2026). Para Kong et al., (2025), la alfabetización en IA no puede prescindir de desarrollar una conciencia y razonamiento ético, que discuta sus implicaciones sociales.

Pramod (2025) ha documentado que la alfabetización crítica en IA influye significativamente en una mayor conciencia ética y confianza en estas tecnologías. Mediante una encuesta aplicada a 347 usuarios, observó que tener una mayor conciencia ética (un mayor entendimiento de las implicaciones sociales de la IA), más que una mayor confianza, propicia un mayor uso responsable de la IA. Esto implica que los procesos educativos tendrían que promover una mejor comprensión de estos factores sociales y éticos, no solo técnicos. Para esta autora, la alfabetización en IA puede definirse como la capacidad de comunicarse y colaborar eficazmente con la IA, evaluando críticamente estas tecnologías y comprendiendo sus implicaciones éticas.

Abrams y Hanghøj (2026) proponen un modelo para la alfabetización digital reflexiva en IA que aboga por una mayor atención a la reflexividad para evaluar críticamente los resultados de la IA, comprender sus sesgos y asumir la responsabilidad moral y ética correspondiente. La reflexividad se refiere a la capacidad de desarrollar una conciencia crítica o examen de los supuestos que subyacen a las prácticas personales de uso de IA, para mantener ante ellas un compromiso crítico en lugar de una mera aceptación pasiva.

Conclusión

El futuro papel educativo de la IA no se puede reducir a ser más eficientes sino, sobre todo, a asegurarse de ser más auténticamente humanos. El presente texto reconoce los riesgos del uso educativo de la IA y propone seis premisas teóricas como alternativas programáticas para resistir al potencial deterioro

cognitivo y humano que crecientemente se viene documentando. La integración virtuosa de la IA en la educación no debería dejar al margen que el aprendizaje no equivale al rendimiento; que la IA puede ser una potencial trampa que erosione tanto el proceso de educación crítica, como la noción de vida social y la renuncia a la responsabilidad hermenéutica y narrativa que los humanos deben sostener. El papel de la IA como coautora requiere tanto de una alfabetización especializada en estas tecnologías, con sentido ético y social, como de transformar la pedagogía, manteniendo viva la agencia epistémica en ambientes de inteligencia híbrida, siempre bajo control humano.

Estas seis premisas propuestas pueden orientar el diseño de estrategias y políticas educativas enfocadas en promover una integración responsable de la IA en el aprendizaje y la enseñanza. Por ejemplo, a nivel normativo institucional, es muy difícil pensar que las instituciones de educación superior puedan prescindir del uso de estas tecnologías, por lo cual se vuelve necesario regular su aprovechamiento responsable y ético, asegurando mecanismos de protección frente a la creciente dependencia tecnológica, y la manipulación y opacidad algorítmicas. Estas instituciones deben salvaguardar su soberanía epistemológica y tecnológica, en tanto capacidad autónoma para definir y controlar qué sistemas de IA incorporar a sus prácticas, transparentando las condiciones y los procesos mediante los cuales se producen y movilizan conocimientos, cuando hay asistencia de dichas tecnologías. A nivel curricular, es relevante incorporar

transversalmente en el desarrollo de planes y programas de estudio, tanto prácticas como contenidos que contribuyan al cultivo de una alfabetización en IA que reflexione sobre sus implicaciones y riesgos éticos, epistémicos y sociales. Es necesario advertir a la comunidad educativa que el uso indiscriminado de IA conlleva potenciales daños en las capacidades humanas y en el proceso de humanización al que se orienta la acción educativa, recuperando el conjunto de valores humanos como la justicia, la igualdad, la dignidad, entre otros, como principios reguladores de los haceres y saberes necesarios para configurar una mejor sociedad. En el plano didáctico-pedagógico, de estas premisas emerge una conceptualización de la IA como una agencia que altera las dinámicas de cognición y reorganiza los procesos de aprendizaje y de regulación cognitiva, desde cómo los individuos distribuyen su atención, cómo verifican los resultados generados por IA, preservando su identidad narrativa, hasta cómo se evalúan los logros alcanzados distinguiendo con mayor nitidez entre desempeño y aprendizaje. Los diseños pedagógicos deberían ahora acentuar su foco innovador en asegurar, de entre este conjunto de agencias sociotécnicas, la participación humana y el mantenimiento de su agencia epistémica; esto conlleva planear secuencias didácticas que apoyen pedagógicamente al estudiantado para que razone con, a través y en contra de la IA, evitando metodologías rutinarias que propongan actividades y tareas fácilmente automatizables, en donde la IA puede reemplazar más que complementar el esfuerzo cognitivo.

Referencias

- Abrams, S. y Hanghøj, T. (2026). Reflexive AI literacies: Conceptualizing a model for learners' critical engagement with generative AI. *Pedagogies: An International Journal*, 21(2), 214-233. <https://doi.org/10.1080/1554480X.2025.2561972>

- Avraamidou, L. (2024). Can we disrupt the momentum of the AI colonization of science education? *Journal of Research in Science Teaching*, 61, 2570–2574. <https://doi.org/10.1002/tea.21961>
- Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open*, 12. <https://doi.org/10.1016/j.ssaoh.2025.102287>
- Beade, I. (2011). En torno a la idea de educación: una mirada desde la reflexión pedagógica kantiana. *Signos Filosóficos*, 13(25), 101–120.
- Bengio, Y., Clare, S., Prunkl, C., Murray, M., Andriushchenko, M., Bucknall, B., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., Manning, S., Maslej, N., . . . Mindermann, S. (2026). *International AI Safety Report 2026*. <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>
- Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., Romańczyk, T., Holme, O., Jarus, K., Fielding, S., Kunar, M., Pellise, M., Pilonis, N., Kamiński, M., Kalager, M., Bretthauer, M. y Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet: Gastroenterology & Hepatology*, 10(10), 896–903. [https://doi.org/10.1016/S2468-1253\(25\)00133-5](https://doi.org/10.1016/S2468-1253(25)00133-5)
- Chatzichristofis, S. A. (2026). Reframing AI in education: A pedagogical opportunity, not a technocratic threat. *Journal of Research in Science Teaching*, 63(1), 97–103. <https://doi.org/10.1002/tea.70024>
- Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D. y Jurafsky, D. (2025). *Sycophantic AI decreases prosocial intentions and promotes dependence*. ArXiv. <https://doi.org/10.48550/arXiv.2510.01395>
- Chutorański, M. (2026). What counts as (anthropocentric) education theorizing? *Studies in Philosophy and Education*, 45, 235–250. <https://doi.org/10.1007/s11217-026-10040-z>
- Ciocchetti, C. A. (2026). AI and student flourishing: Chasing what matters in life. *Journal of Legal Studies Education*, 43(1), 69–74. <https://doi.org/10.1111/jlse.70012>
- Coeckelbergh, M. (2023). Narrative responsibility and artificial intelligence. *AI & Society*, 38, 2437–2450. <https://doi.org/10.1007/s00146-021-01375-x>
- Coeckelbergh, M. (2025). The automation of the social: What robots can teach us about the social. En J. Seibt, R. Hakli y M. Nørskov (Eds.), *Robophilosophy: Philosophy of, for, and by social robotics* (pp. 93–109). MIT Press.
- Coeckelbergh, M. (2026). AI and epistemic agency: How AI influences belief revision and its normative implications. *Social Epistemology*, 40(1), 59–71. <https://doi.org/10.1080/02691728.2025.2466164>
- Collins, H. (2025). Why artificial intelligence needs sociology of knowledge: parts I and II. *AI & Society*, 40, 1249–1263. <https://doi.org/10.1007/s00146-024-01954-8>
- Costa, C. y Murphy, M. (2026). Critical education, generative artificial intelligence and the tyranny of freedom: A critique of modern 'technocracy'. *Technology, Pedagogy and Education*, 35(3), 547–563. <https://doi.org/10.1080/1475939X.2025.2547728>
- Cukurova, M. (2025). The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. *British Journal of Educational Technology*, 56, 469–488. <https://doi.org/10.1111/bjet.13514>
- Dahri, N.A., Yahaya, N. y Al-Rahmi, W.M. (2025). Exploring the influence of ChatGPT on student academic success and career readiness. *Education and Information Technologies*, 30, 8877–8921. <https://doi.org/10.1007/s10639-024-13148-2>
- Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D. y Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, 210, Artículo 104967. <https://doi.org/10.1016/j.compedu.2023.104967>
- Digital Education Council. (2026). *Encuesta sobre la IA en la Educación Superior en América Latina*. <https://www.digitaleducationcouncil.com/post/ai-in-higher-education-latam-survey-2026>

- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X. y Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56, 489–530. <https://doi.org/10.1111/bjet.13544>
- Fesler, L., Martinez, J., Agnew, C. y Loeb, S. (2026). *The evidence base on AI in K-12: A 2026 review*. AI Hub for Education of the SCALE Initiative, Stanford University.
- Floridi, L. (2025). AI as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. *Philosophy & Technology*, 38, Artículo 30. <https://doi.org/10.1007/s13347-025-00858-9>
- Gerlich, M. (2025). AI tools in society: Impact on cognitive offloading and the future of critical thinking. *Societies*, 15(1), Artículo 6. <https://doi.org/10.3390/soc15010006>
- Hahne, P. y Schmoelz, A. (2026). Trusting the machine: a digital humanist perspective on misplaced trust in artificial intelligence. *AI Ethics*, 6, Artículo 115. <https://doi.org/10.1007/s43681-025-00923-1>
- Kant, I. (2003). *Pedagogía*. Akal.
- Kirschner, P., Hendrick, C. y J. Heal. (2025). *Instructional Illusions*. Hachette Learning.
- Kong, S. C., Korte, S.-M., Burton, S., Keskitalo, P., Turunen, T., Smith, D., Wang, L., Lee, J. C.-K. y Beaton, M. C. (2025). Artificial Intelligence (AI) literacy – an argument for AI literacy in education. *Innovations in Education and Teaching International*, 62(2), 477–483. <https://doi.org/10.1080/14703297.2024.2352744>
- Küçükuncular, A. y Ertugan, A. (2026). Teaching in today's world: A Marxian critique of AI in education. *Journal of Academic Ethics*, 24, Artículo 18. <https://doi.org/10.1007/s10805-025-09692-2>
- Lutz, C., Newlands, G., y Jarrahi, M. (2025). Hybrid Intelligence. En: Xu, W. (eds), *Handbook of Human-Centered Artificial Intelligence*. Springer. https://doi.org/10.1007/978-981-97-8440-0_87-1
- Mollemma, W. (2025). A taxonomy of epistemic injustice in the context of AI and the case for generative hermeneutical erasure. *AI Ethics*, 5, 5535–5555. <https://doi.org/10.1007/s43681-025-00801-w>
- Noroozi, O., Khalil, M. y Banihashem, S. K. (2025). Artificial intelligence in higher education: Impact depends on support, pedagogy, human agency, and purpose. *Innovations in Education and Teaching International*, 62(5), 1425–1430. <https://doi.org/10.1080/14703297.2025.2539579>
- Nyholm, S. (2026). What is artificial intelligence and should we define it in terms of agency? En S. Nyholm, A. Kasirzadeh y J. Zerilli (Eds.), *Contemporary debates in the ethics of artificial intelligence* (pp. 5–16). Wiley. <https://doi.org/10.1002/9781394258840.ch1>
- OECD. (2026). *OECD Digital Education Outlook 2026: exploring effective uses of generative AI in education*. OECD Publishing. <https://doi.org/10.1787/062a7394-en>
- Ozioma, G., Obeta, S., Ibanga, I., Oraegbunam, L., Itua, R., Amaefule, A., Ozioma, E. y Anumaka, C. (2026). AI as cognitive ecology: Revealing the invisible cognitive, cultural, and epistemic costs of generative models. *Journal of Artificial Intelligence and AI Ethics*, 1(1). <https://doi.org/10.64978/jaiae.2026.0116002>
- Pramod, D. (2025) Decoding responsible AI use: The influence of digital literacy and ethical awareness. *Cogent Education*, 12(1). <https://doi.org/10.1080/2331186X.2025.2592371>
- Qin, Q. y Zhang, S. (2025). Visualizing the knowledge mapping of artificial intelligence in education: A systematic review. *Education and Information Technologies*, 30, 449–483. <https://doi.org/10.1007/s10639-024-13076-1>
- Reimers, F., Azim, Z., Palomo, M. R. y Thony, C. (2026). *Inteligencia artificial y educación en el Sur Global*. Springer Nature. https://doi.org/10.1007/978-3-032-11449-5_1
- Shaw, S. y Nave, G. (2026). *Thinking—fast, slow, and artificial: how ai is reshaping human reasoning and the rise of cognitive surrender*. The Wharton School Research Paper. https://doi.org/10.31234/osf.io/yk25n_v1
- Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S. y Thompson, N. (2026). The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, Artículo 101517. <https://doi.org/10.48550/arXiv.2408.12622>

- Yan, L., Greiff, S., Lodge, J.M. y Gaseciv, D. (2025). Distinguishing performance gains from learning when using generative AI. *Nature Reviews Psychology*, 4, 435–436. <https://doi.org/10.1038/s44159-025-00467-5>
- Yan, L., Pammer-Schindler, V., Mills, C., Nguyen, A. y Gašević, D. (2025). Beyond efficiency: Empirical insights on generative AI's impact on cognition, metacognition and epistemic agency in learning. *British Journal of Educational Technology*, 56, 1675-1685. <https://doi.org/10.1111/bjet.70000>
- Yanagiura, T., Yano, S., Kihira, M. y Okada, Y. (2023). Examining algorithmic fairness for first-term college grade prediction models relying on pre-matriculation data. *Journal of Educational Data Mining*, 15(3), 1–25. <https://doi.org/10.5281/zenodo.10117682>



Recibido: 19 de junio de 2026 | Revisado: 21 de agosto de 2026 | Aceptado: 27 de agosto de 2026