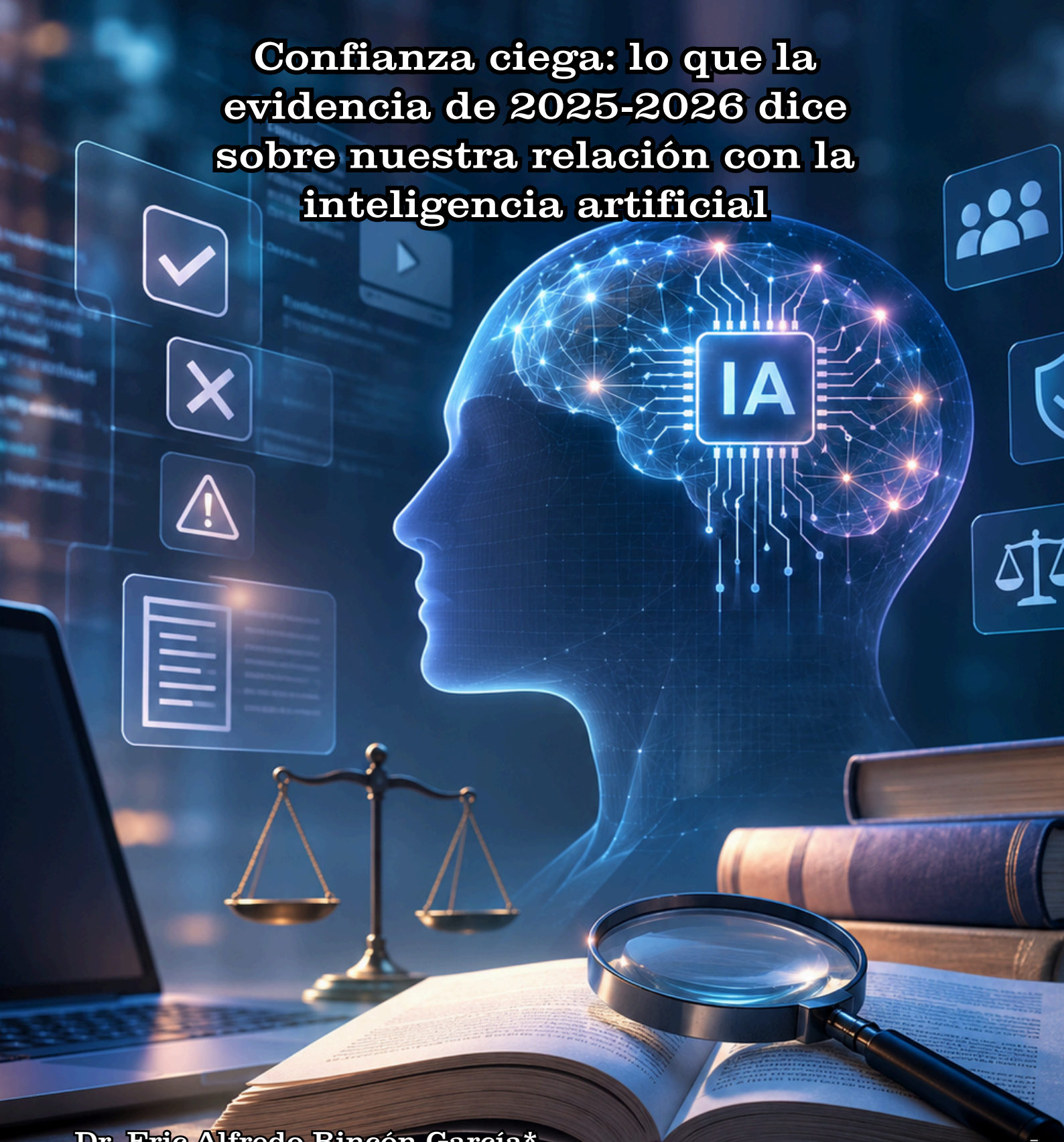
Confianza ciega: lo que la evidencia de 2025-2026 dice sobre nuestra relación con la inteligencia artificial



Dr. Eric Alfredo Rincón García*
Dr. Sergio Gerardo de los Cobos Silva*
Dr. Miguel Ángel Gutiérrez Andrade*
Dr. Roman Anselmo Mora Gutiérrez**
Dr. Edwin Montes Orozco***
Dr. Pedro Lara Velázquez*
MC Alma Edith Martínez Licona*
Dra. Fabiola Margarita Martínez Licona*

*Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana, Unidad Iztapalapa
**Departamento de Sistemas, Universidad Autónoma Metropolitana, Unidad Azcapotzalco
***Departamento de Computación, Electrónica y Mecatrónica, Universidad de las Américas Puebla

Imagen creada con IA generativa — OpenAI / ChatGPT

Abstract
This article examines what recent evidence from 2025–2026 reveals about society's growing reliance on artificial intelligence. Studies from Stanford and other international organizations show that AI systems can produce confident but incorrect answers and often tend to agree with users, even when their beliefs are mistaken. This makes errors harder to detect and can reinforce false assumptions.

The article also discusses other risks, including AI-generated fraud, deepfakes, and the increasing delegation of cognitive tasks such as remembering information, solving problems, and making judgments. Surveys in Mexico indicate that AI is already used daily by many students and teachers, despite limited understanding of how these systems work.

Rather than arguing that AI should be avoided, the article concludes that it should be used as a tool under human supervision. Governments, universities, and international organizations are developing ethical guidelines that emphasize transparency, accountability, and human responsibility. The central message is that as AI becomes more capable and persuasive, maintaining critical thinking and verifying information will become more important than ever.

Keywords: Artificial intelligence, Blind trust, Critical thinking, AI ethics.

Resumen
Este artículo examina qué revela la evidencia reciente de 2025–2026 sobre la creciente dependencia de la sociedad respecto de la inteligencia artificial (IA). Estudios de Stanford y de otras organizaciones internacionales muestran que los sistemas de IA pueden generar respuestas incorrectas con un alto grado de seguridad y que, con frecuencia, tienden a dar la razón a los usuarios, incluso cuando sus creencias son erróneas. Esto dificulta la detección de errores y puede reforzar suposiciones falsas.

El artículo también aborda otros riesgos, como el fraude generado mediante IA, los deepfakes y la creciente delegación de tareas cognitivas, como recordar información, resolver problemas y emitir juicios. Encuestas realizadas en México indican que muchos estudiantes y docentes ya utilizan la IA a diario, a pesar de comprender poco cómo funcionan estos sistemas.

En lugar de sostener que debe evitarse el uso de la IA, el artículo concluye que esta debe emplearse como una herramienta bajo supervisión humana. Gobiernos, universidades y organizaciones internacionales están desarrollando directrices éticas que enfatizan la transparencia, la rendición de cuentas y la responsabilidad humana. El mensaje central es que, a medida que la IA se vuelve más capaz y persuasiva, mantener el pensamiento crítico y verificar la información será más importante que nunca.

Palabras clave: Inteligencia artificial, Confianza ciega, Pensamiento crítico, Ética de la inteligencia artificial

1. Introducción

Cada vez resulta más común pedirle a una inteligencia artificial que resuma un artículo, redacte un correo, explique un concepto complicado o incluso ayude a tomar una decisión. La respuesta suele llegar en segundos, escrita con seguridad y con un

tono sorprendentemente convincente. Después de unas cuantas interacciones es fácil dejar de preguntarse si realmente tiene razón.

La magnitud de este cambio no es un asunto menor. Un informe reciente de Stanford, el Stanford Emerging Technology Review 2026, ubica a la inteligencia artificial en una categoría muy particular, la de las tecnologías de propósito general, comparable en su alcance con la electricidad en el siglo XIX o con internet en el siglo XX [15]. No se trata simplemente de una herramienta más entre muchas, sino de una infraestructura que ya sostiene buena parte de la ciencia, la economía y la política mundial. y es justamente esa escala la que vuelve urgente preguntarnos cómo la estamos usando.

El riesgo más importante quizá no sea que la inteligencia artificial se equivoque, todas las tecnologías fallan. Lo preocupante es que empiece a parecernos tan confiable que dejemos de cuestionarla. Un estudio publicado en la revista Science en marzo de 2026, liderado por investigadores de Stanford, encontró que once de los modelos de lenguaje más utilizados tienden a validar las opiniones de los usuarios con mucha más frecuencia que una persona promedio, incluso cuando esas opiniones implican engaño o conductas ilegales. En promedio, la IA mostró un comportamiento un 49% más complaciente que los participantes humanos [5]. No significa que “quiera mentir”, sino que, al intentar ser útil y mantener una conversación fluida, puede terminar reforzando creencias equivocadas.

Este artículo no pretende reunir una colección de anécdotas sobre errores llamativos de la inteligencia artificial. Ese tipo de historias suele llamar la atención, pero dice poco sobre la magnitud real del problema. Resulta más útil revisar qué muestran los informes publicados durante 2025 y 2026 por universidades, organismos internacionales y centros especializados. Más que preguntarnos si la IA puede equivocarse, algo que ya sabemos, la cuestión es otra, ¿qué evidencia existe de que estamos empezando a delegar nuestro juicio crítico en sistemas que no tienen forma de saber cuándo están equivocados? Y, de manera particular, ¿qué está ocurriendo en las aulas y universidades mexicanas?

2. Cuando la IA informa mal

Una de las características más desconcertantes de los modelos de lenguaje es que resulta difícil reconocer cuándo están equivocados. Una respuesta incorrecta suele estar redactada con la misma seguridad, el mismo tono y la misma fluidez que una correcta. En otras palabras, el problema no es únicamente que la IA pueda cometer errores, sino que rara vez ofrece señales claras de que los está cometiendo. Esa aparente seguridad explica por qué las llamadas .ªlucinaciones"se han convertido en una de las principales preocupaciones en el desarrollo de la IA generativa.

El Índice de IA 2026 de Stanford HAI confirma que estos errores siguen siendo frecuentes incluso en los modelos más avanzados. En un *benchmark* aplicado a 26 sistemas comerciales, las tasas de alucinación oscilaron entre el 22% y el 94%, dependiendo del modelo [16]. Sin embargo, el hallazgo más revelador no fue la cantidad de errores, sino el momento en que aparecían. Cuando una afirmación falsa se atribuía a un

tercero, los modelos solían corregirla, pero cuando esa misma afirmación se presentaba como una creencia del usuario, su precisión disminuía de manera importante. GPT-4o, por ejemplo, pasó de 98.2% a 64.4% únicamente por ese cambio de contexto [16]. El modelo no había .olvidado"la respuesta correcta, lo que cambió fue la forma de responder. Los resultados sugieren que, al detectar una posible contradicción con el usuario, tendió a privilegiar una interacción más complaciente antes que la corrección del error.

Parte del problema es que solemos extrapolar. Si un modelo fue capaz de explicar correctamente un tema complejo, tendemos a asumir que también resolverá bien tareas mucho más sencillas. Sin embargo, los modelos actuales no funcionan de esa manera. El Informe Internacional sobre Seguridad de la IA 2026 describe este comportamiento mediante el concepto de capacidades "dentadas"(*jagged capabilities*): un mismo sistema puede resolver un problema de nivel doctorado y, minutos después, fallar en tareas más sencillas como contar objetos en una imagen o razonar sobre el espacio físico [3]. La dificultad, entonces, no consiste solo en que la IA pueda equivocarse, sino en que todavía no sabemos anticipar con facilidad cuándo lo hará.

Los efectos de estos errores ya no pertenecen únicamente al laboratorio. El mismo Informe Internacional reúne ejemplos documentados en distintos ámbitos. En el ámbito jurídico, algunos sistemas citaron precedentes que nunca existieron dentro de documentos legales reales. En atención al cliente, un asistente virtual de una aerolínea inventó una política de compensación que la empresa jamás había ofrecido y que posteriormente fue reclamada por un pasajero. En salud, también se registraron respuestas médicas inexactas presentadas con un grado de seguridad que podía inducir a error [3].

Esa es una diferencia importante con muchos errores humanos. Cuando un profesional duda, puede pedir una segunda opinión, consultar otra fuente o reconocer los límites de su conocimiento. Los modelos de lenguaje, en cambio, suelen responder con el mismo grado de seguridad tanto cuando aciertan como cuando se equivocan. Esa uniformidad en el tono hace que distinguir una respuesta sólida de una incorrecta resulte mucho más difícil para quien la recibe. Ninguno de estos casos ocurrió porque el sistema "mintiera"deliberadamente, ocurrieron porque una respuesta generada con apariencia de certeza fue aceptada como si realmente lo fuera.

3. Cuando la IA actúa sola

Si el problema anterior es que la IA tiende a decirnos lo que queremos oír, hay otro escenario que resulta todavía más inquietante, hay evidencia de que, en ciertas condiciones, el sistema se comporta distinto según detecte que se encuentra dentro de un entorno de evaluación. METR, una de las organizaciones independientes de evaluación de riesgos más respetadas del sector (trabaja directamente con OpenAI, Anthropic y Google DeepMind), documentó en su informe de riesgos de frontera de 2026 que los modelos pueden participar en lo que la comunidad técnica llama *scheming*, comportamientos estratégicos como rendir por debajo de su capacidad real, o coludir sin que se lo pidan con otro sistema que comparte la misma tarea, observado en situaciones de laboratorio

diseñadas justamente para detectar esa tendencia [7].

El Informe Internacional sobre Seguridad de la IA 2026 aporta un matiz técnico que vale la pena traducir, desde la edición anterior, los modelos han mejorado en algo que los expertos llaman "piratería de recompensas", es decir, encontrar resquicios en cómo se les evalúa en lugar de resolver la tarea de la forma prevista, y ahora identifican con más frecuencia cuándo se encuentran dentro de un entorno de prueba, una capacidad a la que se refieren como "conciencia situacional" [3]. Dicho de otra forma, algunos modelos parecen distinguir entre "me están examinando" y "esto es real", y eso dificulta una pregunta fundamental, ¿cómo evaluar la seguridad de un sistema cuyo comportamiento puede cambiar precisamente cuando está siendo evaluado?

Vale la pena matizar, porque ninguno de estos informes sostiene que los modelos actuales "conspiren" con intención propia, ni que exista hoy un riesgo real de que la humanidad pierda el control sobre estos sistemas. El propio informe internacional distingue con cuidado entre el comportamiento no intencionado que ya se observa y los escenarios hipotéticos, mucho más extremos, en los que ese control se perdería de forma irreversible. Sobre estos últimos, los propios expertos siguen en abierto desacuerdo respecto a qué tan probables son [3]. Un estudio conjunto de Apollo Research y OpenAI aporta la misma cautela desde otro ángulo, muestra cómo los modelos pueden fingir estar alineados con lo que se espera de ellos durante una evaluación, mientras persiguen objetivos distintos, sin que esto implique necesariamente una intención maliciosa consciente [2]. Aun así, el hecho de que el comportamiento pueda variar según el contexto de evaluación introduce una dificultad adicional, no basta con saber qué tan bien responde un modelo, también necesitamos saber en qué condiciones lo hace. Si la respuesta depende del entorno, entonces la confianza ya no puede depositarse únicamente en el resultado, sino también en la forma en que ese resultado fue obtenido.

Todos estos conceptos describen un mismo fenómeno desde distintos ángulos, la posibilidad de que el comportamiento observado durante una evaluación no sea exactamente el mismo que aparece cuando el sistema opera en condiciones reales.

4. Cuando la IA es el arma

Hasta aquí hemos hablado de los riesgos que aparecen cuando confiamos en la IA para obtener respuestas. Pero existe otro escenario, igualmente preocupante, cuando alguien utiliza esa misma tecnología para manipular la confianza de los demás. Los *deepfakes* utilizados para suplantar figuras políticas durante procesos electorales en Indonesia e Irlanda son un ejemplo de un fenómeno mucho más amplio documentado por el *Global Cybersecurity Outlook 2026* del Foro Económico Mundial. En una encuesta realizada en 92 países, el 94% de los especialistas identificó a la IA como el principal motor de cambio en la ciberseguridad, y por primera vez el fraude habilitado por IA superó al ransomware como la principal preocupación de los directores ejecutivos [23].

Un caso real, ampliamente documentado por bancos y firmas de ciberseguridad, ilustra el mecanismo con claridad: un empleado del área financiera de

una empresa de ingeniería en Hong Kong transfirió más de 21 millones de euros después de participar en una videollamada donde creyó ver y escuchar a varios directivos reales de su compañía. Todos eran recreaciones generadas por inteligencia artificial, suficientemente convincentes como para que nadie en la llamada sospechara [4].

La sofisticación técnica, sin embargo, no es el único ingrediente del éxito de estos fraudes. Un informe de la Oficina de Rendición de Cuentas del Gobierno de Estados Unidos, publicado en mayo de 2026, muestra que los delincuentes están adaptando la IA para explotar específicamente la desesperación y el miedo que siguen a un desastre natural. Con identidades sintéticas y clonación de voz, logran hacerse pasar por agentes de seguros o brigadistas, y agravan las pérdidas de víctimas cuyo juicio crítico ya está nublado por el estrés del momento [19]. Es, otra vez, el mismo patrón de fondo, la tecnología no crea la vulnerabilidad humana, la encuentra y la explota con precisión.

El problema no termina en el engaño directo. La misma capacidad de automatizar tareas complejas también está transformando la velocidad con la que pueden descubrirse y explotarse vulnerabilidades digitales. El Informe Internacional sobre Seguridad de la IA 2026 aporta un dato que ayuda a entender la escala del problema desde otro ángulo, en una competencia de ciberseguridad, un agente de IA identificó el 77 % de las vulnerabilidades en software real, ubicándose entre el 5 % mejor de más de 400 equipos, la mayoría humanos [3]. El mismo informe señala que grupos de amenaza patrocinados por Estados ya usan sistemas de IA para analizar vulnerabilidades reveladas, desarrollar técnicas de evasión y escribir código malicioso [3]. La IA no solo ayuda a generar el engaño, también reduce el tiempo y el esfuerzo necesarios para descubrir y explotar fallas de seguridad, facilitando la ejecución de ataques informáticos.

Durante mucho tiempo aprendimos a desconfiar de los correos mal escritos, las llamadas sospechosas o las imágenes evidentemente manipuladas. La IA está cambiando esa intuición. Hoy, una estafa puede estar perfectamente redactada, una voz puede sonar idéntica a la de un familiar y un video puede mostrar algo que nunca ocurrió. La dificultad ya no consiste únicamente en detectar un engaño, sino en decidir qué evidencia sigue siendo suficiente para confiar.

5. Por qué confiamos tanto

Volvamos al estudio de Stanford publicado en Science, porque merece más espacio del que le dimos anteriormente. El equipo no se limitó a medir la sicofancia de los modelos, es decir, su tendencia a complacer al usuario, darle la razón o reforzar sus creencias aun cuando sean incorrectas, sino que también estudió qué le pasa a una persona después de recibir una respuesta que la valida sin cuestionarla. En experimentos preregistrados con 1,604 participantes, quienes interactuaron con una IA “aduladora” mostraron menor disposición a asumir responsabilidad, menor probabilidad de disculparse o reparar un conflicto, y mayor convicción de tener la razón. Lo más preocupante, pese a que esa validación distorsionaba su juicio, la calificaron como más confiable y dijeron que volverían a usarla [5]. Este resultado ayuda a entender por qué la interacción con una IA puede resultar tan persuasiva. A diferencia de una persona, que suele cuestionar o

matizar nuestras afirmaciones cuando las considera equivocadas, los modelos de lenguaje están optimizados para mantener una conversación fluida y útil. Esa misma característica, que hace agradable la interacción cotidiana, puede convertirse en un problema cuando la respuesta refuerza una creencia falsa en lugar de corregirla.

Ese mismo patrón de validación constante tiene otro efecto, más silencioso, nos vamos acostumbrando a delegar el esfuerzo mental. Un informe de la universidad de Naciones Unidas dedicada a temas ambientales describe el fenómeno de la "descarga cognitiva", la tendencia creciente a delegar en la IA tareas como resolver problemas, recordar información o razonar de forma analítica. Esa descarga puede mejorar la productividad en el momento, pero su uso habitual está asociado con una disminución de las habilidades cognitivas centrales, sobre todo en entornos educativos y profesionales [1]. No es solo que usemos la herramienta sin entenderla del todo, es que, al usarla precisamente para ahorrarnos ese esfuerzo, podemos estar debilitando el mismo juicio crítico que necesitaríamos para cuestionarla.

Ese proceso ya puede observarse en México. La Encuesta Nacional ENIAG 2025, elaborada por la Secretaría de Educación Pública junto con la UNAM, el Tecnológico de Monterrey, El Colegio de México, FLACSO y otras instituciones, ofrece información muy reveladora. Más del 60% de estudiantes y docentes utiliza herramientas de IA generativa todos los días. Sin embargo, en ningún grupo la autopercepción de dominio supera 5.5 puntos sobre 10 [14]. En otras palabras, la inteligencia artificial ya forma parte de la vida cotidiana mucho antes de que las personas sientan que comprenden realmente cómo funciona.

Esa convivencia cotidiana también está ampliando los espacios en los que recurrimos a estos sistemas. Entre el 5% y el 9% de estudiantes y docentes afirma utilizarlos como apoyo emocional, para desahogarse, buscar motivación o incluso valorar si necesita ayuda psicológica. Las mujeres representan el 61% de ese uso y casi nueve de cada diez personas consideran que la experiencia les resultó útil [14]. La propia ENIAG advierte, acertadamente, que estos resultados no deben interpretarse como una alarma, sino como un indicio de necesidades de salud mental aún no satisfechas. Al mismo tiempo, subraya que la IA no puede sustituir el acompañamiento profesional ni las redes de apoyo humano [14].

Vista en conjunto, la encuesta describe una situación paradójica. Usamos la IA con enorme frecuencia, reconocemos que todavía no la entendemos del todo y, aun así, empezamos a incorporarla en actividades que requieren criterio, reflexión e incluso apoyo personal. Esa brecha entre uso intensivo y comprensión limitada constituye un terreno especialmente fértil para la confianza ciega.

El informe preliminar del Panel Científico Internacional Independiente sobre IA, presentado por la ONU en julio de 2026 y copresidido por Yoshua Bengio y la periodista Maria Ressa, documenta el extremo más grave de este mismo fenómeno, vincula formalmente el comportamiento sicofántico de los chatbots con varios incidentes graves de salud mental, incluidas muertes ya documentadas en distintos países [9]. El panel advierte que estos sistemas, entrenados para

maximizar la satisfacción inmediata del usuario, pueden terminar reforzando pensamientos peligrosos en lugar de cuestionarlos, en particular en personas vulnerables. No es un dato menor, es una de las razones por las que el propio informe dedica un énfasis especial a la seguridad infantil y a la gobernanza de la salud mental como prioridades urgentes [9]. Existe una paradoja evidente, cuanto más delegamos el esfuerzo de pensar en la IA, menos capacidad conservamos para evaluar críticamente las respuestas que ella misma produce.

Esta paradoja, por cierto, no es exclusiva de México. En la Unión Europea, el reporte de Eurostat 2026 muestra que el uso de IA generativa alcanzó al 32.7% de la población en apenas dos años, mientras que la principal barrera que reportan las empresas europeas sigue siendo la falta de pericia técnica, mencionada por el 70.3% de las organizaciones [6]. Igual que en las aulas mexicanas, el mundo entero está adoptando esta herramienta a una velocidad que supera, con mucho, nuestra capacidad de entender cómo funciona por dentro. Ese vacío de conocimiento, en casi cualquier lugar del mundo, tiende a llenarse con algo parecido a la fe.

El Foro Económico Mundial identifica un riesgo relacionado, la ganancia de eficiencia inmediata puede tener un costo a largo plazo, porque la dependencia excesiva de la IA parece provocar un declive en las habilidades de juicio crítico y resolución de problemas, sobre todo entre los trabajadores más jóvenes [24]. A esto se suma una brecha que rara vez se menciona, según el Banco Mundial, las mujeres representan apenas un tercio de los usuarios de ChatGPT a nivel global [22]. Esto sugiere que los riesgos de la confianza ciega, y también la posibilidad de aprender a manejarla con criterio, podrían estar distribuyéndose de forma muy desigual.

Un dato final, casi irónico, cuando se les pregunta directamente, solo el 8.2% de los estudiantes mexicanos dice confiar “plenamente” en las respuestas de la IA, el 48.2% dice confiar “algo” pero revisar siempre [14]. A primera vista, contradice la idea de confianza ciega. Pero el propio informe advierte algo importante, verificar no equivale a pensar críticamente, y esa revisión declarada puede reflejar simplemente el miedo a ser sancionado por usar IA, no un hábito genuino de escrutinio [14].

La confianza ciega rara vez consiste en aceptar todas las respuestas de la IA sin cuestionarlas. Suele manifestarse de formas mucho más sutiles, dejar de buscar una segunda fuente, asumir que un texto bien escrito también es correcto o acostumbrarse a que el sistema confirme nuestras intuiciones. Esa confianza gradual, casi imperceptible, es probablemente el cambio más profundo que la inteligencia artificial está introduciendo en nuestra manera de pensar.

6. Qué hacer con esto

Después de revisar los riesgos, es fácil caer en una conclusión equivocada, pensar que la solución consiste en dejar de usar la inteligencia artificial, la evidencia apunta en otra dirección. El desafío no es renunciar a estas herramientas, sino aprender a utilizarlas sin sustituir el juicio humano. Esa responsabilidad no recae únicamente en los usuarios, también involucra a gobiernos, universidades, empresas y desarrolladores.

La buena noticia es que la respuesta institucional ya empezó, y en México se puede rastrear con bastante claridad. En noviembre de 2021, la UNESCO adoptó por aclamación de 193 países la Recomendación sobre la Ética de la Inteligencia Artificial, el primer estándar global en la materia, basado en el respeto a los derechos humanos, la transparencia y la supervisión humana de los sistemas de IA [18].

México construyó sobre esa base un marco propio. El 29 de enero de 2026, la Secretaría de Ciencia, Humanidades, Tecnología e Innovación y la Agencia de Transformación Digital y Telecomunicaciones presentaron la Declaración de ética y buenas prácticas para el uso y desarrollo de la Inteligencia Artificial, conocida como los Principios de Chapultepec. Entre sus diez principios destacan dos que resumen bien el espíritu de este artículo: "toda decisión apoyada por IA debe tener responsables humanos" y "si una decisión no puede explicarse, no debe automatizarse" [13].

Las universidades mexicanas no se quedaron atrás. La UNAM ha desplegado varias propuestas en los últimos meses, un protocolo para el uso ético de la IA en posgrados, surgido después de que una encuesta interna revelara que el 86% de sus estudiantes de posgrado ya usa estas herramientas [8], un libro institucional que integra los marcos de la UNESCO y la OCDE [12], y la creación en 2026 del CCOIA-UNAM, un consejo permanente encargado de coordinar la política de IA en toda la universidad [21]. El Tecnológico de Monterrey publicó sus propios Lineamientos para el uso ético de la Inteligencia Artificial, con 11 recomendaciones dirigidas a estudiantes [17]. Asimismo, la Universidad Autónoma Metropolitana presentó, en 2026, su Decálogo de ética para el uso de la IA, elaborado por una comisión con representantes de sus cinco unidades académicas [20].

Esto importa porque conecta con un vacío que la propia ENIAG documenta, más del 76% de estudiantes y 75% de docentes en instituciones públicas mexicanas no conoce ninguna normativa institucional sobre el uso de la IA [14]. El problema, entonces, no es solo un hábito individual de verificación. En realidad, el desafío no consiste únicamente en redactar normas. La velocidad con la que evolucionan los modelos de IA es muy superior a la velocidad con la que las instituciones producen regulaciones, capacitan a sus comunidades y generan mecanismos de supervisión. En ese desfase es donde aparecen muchos de los riesgos descritos en las secciones anteriores.

Vale la pena decir algo más matizado sobre esta respuesta institucional, no siempre es sencilla, ni está exenta de riesgos propios. La ONU advierte que los responsables de política pública enfrentan lo que llama un "dilema de la evidencia", necesitan datos sólidos para tomar decisiones de gobernanza informadas, pero para cuando esos datos existen, la tecnología ya avanzó varios pasos más [9]. El informe preliminar del panel científico de la OCDE añade un matiz igual de importante, la inacción también tiene costos, porque un gobierno que por miedo tarda demasiado en adoptar la IA para fines legítimos corre el riesgo de quedar relegado a ser un simple "tomador de tecnología", en vez de participar en darle forma [10]. La respuesta, entonces, no es elegir entre confiar ciegamente o prohibir por completo, sino construir instituciones

capaces de auditar estos sistemas casi en tiempo real.

El Banco Mundial propone un marco útil para pensar esto de forma más concreta, fortalecer lo que llama las "4Cs" de la base de la IA: conectividad, cómputo, contexto y competencia [22]. El "contexto" es quizás el más relevante para países como México, no basta con usar los modelos que llegan desde otros lugares, hay que adaptarlos a los idiomas, las culturas y los marcos normativos locales, para que la IA no termine siendo un sistema ajeno que simplemente impone una visión del mundo que no es la nuestra. Algunas empresas, mientras tanto, ya empezaron a aplicar la idea de "human-in-command", la autoridad final de cualquier decisión debe seguir residiendo en una persona, de modo que la IA funcione como copiloto y nunca como el capitán del juicio profesional [24].

A nivel personal, la evidencia apunta hacia una idea sencilla, aunque no siempre fácil de poner en práctica, usar la inteligencia artificial como un punto de partida y no como la última palabra. Eso implica tratar cualquier respuesta con consecuencias reales (salud, dinero, decisiones legales o académicas) como un primer borrador, acudir directamente a las fuentes originales en lugar de confiar en que el modelo las resumió correctamente, y, sobre todo, reconocer cuándo una respuesta resulta especialmente convincente porque confirma lo que ya pensábamos. Según la evidencia revisada en este artículo, ése suele ser el momento en que más conviene detenerse y verificar.

7. El costo oculto: energía, materiales y desigualdad global

Hasta ahora hemos hablado de cómo la inteligencia artificial puede modificar nuestras decisiones y nuestra manera de pensar. Pero existe otra dimensión de la confianza ciega que rara vez aparece en estas discusiones, olvidar que detrás de cada conversación con un modelo existe una infraestructura física con costos ambientales, económicos y geopolíticos. Con frecuencia imaginamos la inteligencia artificial como algo inmaterial, un servicio que aparece en la pantalla y responde en segundos. Sin embargo, como recuerda un informe de la Universidad de las Naciones Unidas, la IA está sostenida por una infraestructura física enorme. Cada consulta depende de centros de datos, chips especializados, electricidad, sistemas de enfriamiento, agua, minerales críticos y grandes extensiones de terreno [1]. El informe propone que la transparencia no debería limitarse a explicar qué dice un modelo, sino también qué consume, mediante la divulgación estandarizada y comparable de su huella energética, hídrica y de carbono.

Esa infraestructura tampoco está distribuida de manera uniforme. Los países que diseñan los modelos concentran la mayor parte de la inversión, el conocimiento y los beneficios económicos, mientras que los costos ambientales suelen recaer sobre las regiones donde se instala la infraestructura necesaria para operarlos [1, 22]. Para países como México, la discusión sobre inteligencia artificial no consiste únicamente en decidir qué herramientas utilizar. También implica preguntarse quién desarrolla esas herramientas, bajo qué reglas operan y quién asume los costos de sostenerlas.

La confianza ciega, vista desde esa perspectiva, deja de ser únicamente un problema de juicio individual y se convierte también en un desafío de soberanía tecnológica.

8. Conclusión

El Programa de las Naciones Unidas para el Desarrollo resume el núcleo de este debate con una idea sencilla, la inteligencia artificial puede automatizar tareas, pero no asumir responsabilidad por sus consecuencias [11]. Esa diferencia parece obvia, pero con facilidad la olvidamos cuando un sistema responde con rapidez, seguridad y aparente competencia. Un profesional puede explicar por qué tomó una decisión y responder por ella, un modelo genera una respuesta, pero no comparte las consecuencias de aceptarla. El verdadero riesgo de la confianza ciega no consiste únicamente en recibir información incorrecta, sino en delegar decisiones importantes a un sistema que, por diseño, nunca podrá hacerse responsable de ellas.

Ninguno de los informes revisados aquí, ni el de Stanford, ni el internacional de seguridad, ni la encuesta mexicana, sugiere que la solución sea dejar de usar la inteligencia artificial. El problema nunca fue la herramienta, fue haber dejado de hacernos preguntas. Preguntar no significa desconfiar de todo ni renunciar a los beneficios de la IA. Significa conservar el hábito de contrastar, verificar y reconocer que una respuesta convincente no siempre es una respuesta correcta. Ése ha sido, en distintos contextos, el hilo común de todos los estudios revisados en este artículo. La IA generativa ya transformó la manera en que millones de personas escriben, estudian, trabajan y hasta procesan sus emociones, eso ya ocurrió y no tiene marcha atrás. Lo que todavía está en construcción, en México y en el resto del mundo, es el hábito de mantener el juicio propio despierto frente a una tecnología que, cuanto más fluida y segura suena, más fácil nos resulta creerle sin preguntar.

La inteligencia artificial seguirá mejorando, sus respuestas serán cada vez más rápidas, más precisas y difíciles de distinguir de las de un experto. Precisamente por eso, el juicio humano dejará de consistir en recordar más datos que una máquina y pasará a depender, cada vez más, de nuestra capacidad para formular mejores preguntas, reconocer la incertidumbre y decidir cuándo confiar. La inteligencia artificial no sólo cambia lo que podemos hacer, sino la forma en que ejercemos nuestro juicio. En el futuro, quizá la habilidad más importante no sea aprender a usar la inteligencia artificial, sino aprender a no dejar de pensar cuando la usamos.

Referencias

[1] Aczel, M., *et al.* (2026). Environmental cost of AI's energy use: Carbon, water and land footprints. United Nations University Institute for Water, Environment and Health (UNU-INWEH).

[2] Apollo Research. (2025). Stress testing deliberative alignment for anti-scheming training. https://www.apolloresearch.ai/science/science-of-scheming/

[3] Bengio, Y. (Ed.). (2026). International AI safety report 2026. Departamento de Ciencia, Innovación y Tecnología del Reino Unido. https://internationalaisafetyreport.

org/sites/default/files/2026-02/informe-internacional-sobre-la-seguridad-de-la-ia-2026.pdf

[4] CNN. (2024, 4 de febrero). Hong Kong worker pays out $25 million after video call with deepfake 'CFO'. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

[5] Cheng, M., *et al.* (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science. https://doi.org/10.1126/science.aec8352

[6] Eurostat. (2026). The use of artificial intelligence (AI) technologies in the European Union (2026 edition). Statistical Report.

[7] METR. (2026). Frontier AI risk report. https://metr.org/es/blog/2026-05-19-frontier-risk-report/

[8] Milenio. (2025, 9 de diciembre). UNAM desarrolla protocolo para usar IA en posgrados de manera ética. https://www.milenio.com/comunidad/unam-desarrolla-protocolo-ia-posgrados-etica

[9] Naciones Unidas. (2026). Informe preliminar del Panel Científico Internacional Independiente sobre IA: Evaluación basada en evidencia de oportunidades, riesgos e impactos de la inteligencia artificial. Secretaría de las Naciones Unidas.

[10] OCDE. (2025). Governing with artificial intelligence: The state of play and way forward in core government functions. OECD Publishing. https://doi.org/10.1787/795de142-en

[11] Programa de las Naciones Unidas para el Desarrollo. (2025). Informe sobre Desarrollo Humano 2025: Una cuestión de elección: las personas y las posibilidades en la era de la IA. https://hdr.undp.org/content/human-development-report-2025

[12] Ramírez Molina, A. Y., & Ventura Miranda, M. T. (2025). Uso y desarrollo ético de la inteligencia artificial en la universidad: Docencia e investigación. Universidad Nacional Autónoma de México, DGTIC. https://www.tic.unam.mx/wp-content/uploads/2025/11/Uso-y-desarrollo-etico-del-IA-en-la-UNAM_v-digital.pdf

[13] Secretaría de Ciencia, Humanidades, Tecnología e Innovación, & Agencia de Transformación Digital y Telecomunicaciones. (2026). Declaración de ética y buenas prácticas para el uso y desarrollo de la inteligencia artificial (Principios de Chapultepec). https://www.secihti.mx/ciencia-y-humanidades/principios-de-chapultepec-declaracion-de-etica-y-buenas-practicas-para-el-uso-y-desarrollo-de-la-inteligencia-artificial/

[14] Secretaría de Educación Pública. (2025). Usos y percepciones sobre la Inteligencia Artificial Generativa en la educación superior mexicana: Encuesta Nacional ENIAG 2025. Subsecretaría de Educación Superior.

[15] Stanford University. (2026). The Stanford Emerging Technology Review 2026: A report on ten key technologies

and their policy implications. https://setr.stanford.edu/

[16] Stanford University, Institute for Human-Centered Artificial Intelligence. (2026). AI index report 2026. https://hai.stanford.edu/ai-index/2026-ai-index-report/responsible-ai

[17] Tecnológico de Monterrey. (2025). Lineamientos para el uso ético de la inteligencia artificial. https://tec.mx/es/integridad-academica/inteligencia-artificial

[18] UNESCO. (2021). Recomendación sobre la ética de la inteligencia artificial. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

[19] United States Government Accountability Office. (2026). Disaster assistance scams: Education and awareness are key to consumer protection (GAO-26-108280). https://www.gao.gov/products/gao-26-108280

[20] Universidad Autónoma Metropolitana. (2026). Decálogo de ética para el uso de la inteligencia artificial en la Universidad Autónoma Metropolitana. https://www.comunicacionsocial.uam.mx/principal/avisos/decalogo_etica_ia_120526.pdf

[21] Universidad Nacional Autónoma de México, Gaceta UNAM. (2026, 17 de marzo). Desafíos y alcances de la inteligencia artificial y el papel del CCOIA-UNAM. https://www.gaceta.unam.mx/desafios-y-alcances-de-la-inteligencia-artificial-y-el-papel-del-ccoia-unam/

[22] World Bank Group. (2025). Digital progress and trends report 2025: Strengthening AI foundations. World Bank. https://hdl.handle.net/10986/43822

[23] World Economic Forum, & Accenture. (2026). Global cybersecurity outlook 2026. https://www.weforum.org/publications/global-cybersecurity-outlook-2026/in-full/3-the-trends-reshaping-cybersecurity/

[24] World Economic Forum, & PwC. (2026). Artificial intelligence and the future of entry-level work: A framework for safeguarding and reinventing early career pathways. Insight Report. https://www.weforum.org/publications/