revista SIGNO Y PENSAMIENTO
Vol. 44
Publicación continua
2025
ISSN: 2027-2731 (En línea) | ISSN: 0120-4823 (Impreso - Descontinuado)

Artículos

Tres reflexiones éticas y prácticas sobre cómo usar la inteligencia artificial en educación [*]

Three Ethical and Practical Reflections on how to Use Artificial Intelligence in Education

Três reflexões éticas e práticas sobre como usar a inteligência artificial na educação

Manuel Enrique Cardoso [a]
Universidad ORT, Uruguay
cardoso@ort.edu.uy
ORCID: https://orcid.org/0000-0002-2706-9839

DOI: https://doi.org/10.11144/Javeriana.syp44.trep

Recibido: 18 julio 2025
Aceptado: 30 septiembre 2025
Publicado: 30 diciembre 2025

Resumen:

Se proponen, desde una perspectiva ética y práctica, tres reflexiones sobre el uso de la inteligencia artificial en educación. Estos debates son iteraciones actualizadas de discusiones antiquísimas, revitalizadas por esta disrupción. La primera reflexión procura proteger a los usuarios de la inteligencia artificial generativa (IAG), en especial a individuos en formación; se inspira en la máxima hipocrática *primum non nocere* ('lo primero es no hacer daño'). Una segunda reflexión explora las contradicciones del poder pedagógico: cuando la IAG comercial irrumpe en 2022, se intenta prohibir que los estudiantes la usen en evaluaciones sumativas, pero sin que esas restricciones se apliquen a quienes enseñamos e investigamos. Esta reflexión aduce, con Séneca y los estoicos, que la filosofía trata de cosas, no de palabras (*philosophia non in verbis sed in rebus est,* luego vulgarizada como *res, non verba)*. La tercera reflexión explora el ideal de medir los aprendizajes que se valoran, o *measure what you treasure (*y no viceversa), a través del ejemplo de las evaluaciones estandarizadas. El riesgo es que las fortalezas y limitaciones de la corrección automática determinen qué competencias se evalúan y cuáles no. Veremos cómo argumentos de efectividad, eficiencia y equidad son nuevamente utilizados a favor de la estandarización en detrimento de la empatía. Estas reflexiones no procuran prohibir ni mandatar la tecnología, sino identificar usos éticos. Para mostrar la permanencia de estos debates, se los vincula con temas recurrentes de nuestras culturas.

Palabras clave: inteligencia artificial, educación, ética, práctica docente, investigación.

Abstract:

From an ethical and practical perspective, three reflections on the use of artificial intelligence in education are proposed. These debates are updated iterations of age-old discussions, revitalized by this disruption. The first reflection seeks to protect users of generative artificial intelligence (GAI), especially individuals in training; it is inspired by the Hippocratic maxim primum non nocere ('first, do no harm'). A second reflection explores the contradictions of pedagogical power: When commercial GAI bursts onto the scene in 2022, attempts are made to prohibit students from using it in summative assessments, but without these restrictions applying to those of us who teach and conduct research. This reflection argues, with Seneca and the Stoics, that philosophy deals with things, not words (philosophia non in verbis sed in rebus est, later vulgarized as res, non verba). The third reflection explores the ideal of measuring what you value, or measure what you treasure (and not vice versa), through the example of standardized assessments. The risk is that the strengths and limitations of automatic correction determine which skills are assessed and which are not; we will see how arguments of effectiveness, efficiency, and equity are once again used in favor of standardization, to the detriment of empathy. These reflections do not seek to prohibit or mandate technology, but rather to identify ethical uses. To show the permanence of these debates, they are linked to recurring themes in our cultures.

Keywords: Artificial Intelligence, Education, Ethics, Teaching Practice, Research.

Resumo:

Três reflexões sobre o uso da inteligência artificial na educação são propostas a partir de uma perspectiva ética e prática. Esses debates são iterações atualizadas de discussões muito antigas, revitalizadas por essa ruptura. A primeira reflexão busca proteger os usuários da inteligência artificial generativa (IAG), especialmente os indivíduos em treinamento; ela se baseia na máxima hipocrática *primum non nocere* ('primeiro não causar dano'). Uma segunda reflexão explora as contradições do poder pedagógico: quando a IAG comercial entra em cena em 2022, são feitas tentativas de proibir os alunos de usá-la em avaliações somativas, mas sem que essas restrições sejam aplicadas àqueles que ensinam e pesquisam. Esta reflexão argumenta, com Sêneca e os estoicos, que a filosofia trata de coisas, não de palavras (*philosophia non in verbis sed in rebus est,* mais tarde vulgarizada como *res, non verba*). A terceira reflexão

Notas de autor

[a] Autor de correspondencia. Correo electrónico: cardoso@ort.edu.uy

explora o ideal de medir o aprendizado que é valorizado, ou medir o que você valoriza (e não vice-versa), por meio do exemplo das avaliações padronizadas. O risco é que os pontos fortes e as limitações da correção automática determinem quais competências são avaliadas e quais não são; veremos como os argumentos de eficácia, eficiência e equidade são novamente usados a favor da padronização, em detrimento da empatia. Essas reflexões não buscam proibir ou exigir tecnologia, mas identificar usos éticos. Para mostrar a permanência desses debates, eles estão vinculados a temas recorrentes em nossas culturas.

Palavras-chave: inteligência artificial, educação, ética, prática docente, pesquisa.

Recientemente, en un congreso, me invitaron a comentar ponencias de colegas sobre el impacto de la inteligencia artificial en el ámbito educativo. Hubo múltiples puntos de vista representados, pero dos de ellos me parecieron particularmente sintomáticos de un debate que estamos viviendo a escala global. Por un lado, un colega de un banco de desarrollo compartió una perspectiva, con la que tuve algunos puntos de acuerdo, que veía a la inteligencia artificial (IA) como la solución a muchos de los problemas de los sistemas educativos, especialmente en países en desarrollo. Por otro lado, un colega de otra organización internacional presentó una visión sustancialmente más crítica y negativa de la IA, con la cual también tuve algunos puntos de coincidencia, pero que parecía adoptar el supuesto de que podemos parar el avance de la IA. Estas dos visiones habrían sido tildadas por Umberto Eco de *integrada* y *apocalíptica*, respectivamente (Eco, 1994).

Llegado mi turno, adopté una postura más matizada. Para mí, la IA no es una panacea para los sistemas educativos. Sucesivas olas de la revolución digital en educación fueron saludadas como la solución a todos nuestros problemas, pero las estadísticas muestran que, hasta ahora, la introducción de avances tecnológicos no solamente no nos ha permitido superar nuestros desafíos, sino que en muchos casos ha introducido nuevas fuentes de inequidad (Bell *et al*., 2020). Por su lado, la historia nos enseña como los luditas fracasaron en su intento de obstruir el avance de la segunda Revolución Industrial, y esto nos obliga a asumir una postura más realista, aunque algunos académicos proponen abierta y explícitamente revivir el movimiento ludita en reacción al avance de la IA (Logan, 2024; Nichols *et al*., 2025).

En este contexto, este artículo propone tres reflexiones éticas, así como prácticas, sobre el uso de la inteligencia artificial en educación. Las dos últimas son más específicas del campo de la educación, mientras que la primera ("lo primero es no hacer daño") es más general, pero es necesario comenzar por ella en el campo de la educación, como en todos los otros. A su vez, estas tres reflexiones no agotan, ni mucho menos, los dilemas éticos con relación al uso de la IA en el campo de la educación. Por ejemplo, dilemas éticos más generales, como los de la propiedad intelectual, la privacidad de los datos de los usuarios y el uso energético de la IA, se aplican tanto al campo de la educación como a todos los demás. Y algunos dilemas más específicos del uso en educación, como la utilización de la IA en evaluaciones de aula (ya sea por parte de los docentes o de los estudiantes), llevarían a discusiones pedagógicas demasiado especializadas como para este espacio académico y disciplinario.

Propongo, entonces, concentrarse en tres aspectos. El primero es la protección de los usuarios de la inteligencia artificial a través de la reflexión en torno a la máxima hipocrática *primum non docere* ('lo primero es no hacer daño'). En segundo lugar, la necesidad de predicar con el ejemplo y modelar buenos usos de la IA para los estudiantes, desde nuestra actividad en la docencia y la investigación, a través del principio *res, non verba* ('hechos y no palabras'), inspirado por los estoicos. Por último, una advertencia sobre el riesgo de que la IA dictamine la agenda de los procesos de evaluación a gran escala; en lugar de esto, debemos procurar medir lo que valoramos (*measure what we treasure*) y no valorar lo que medimos (Cobo, 2016).

Lo primero es no hacer daño: *primum non nocere*

Un principio hipocrático fundamental, luego popularizado en su traducción latina (*primum non nocere*, atribuido a Hipócrates), es que, en el ejercicio de la medicina, lo primero es no hacer daño al paciente, idea que se ha difundido en otras disciplinas. Específicamente, en relación con la inteligencia artificial generativa (IAG), me preocupaban, y aún me preocupan, los riesgos que pueden emerger de la confluencia de tres factores que por separado pueden parecer relativamente inocuos, pero que, cuando actúan de manera conjunta, pueden resultar una combinación explosiva. Esos factores son 1) el sesgo de acuerdo o aquiescencia de la IAG; 2) su tendencia a alucinar, y 3) nuestro propio sesgo de confirmación, combinado con nuestra pereza cognitiva y con las inequidades sociales, económicas y culturales que la IAG puede contribuir a ampliar (Gerlich, 2025).

El sesgo de aquiescencia

Desde el año pasado, he estado usando una versión paga de una herramienta de inteligencia artificial generativa (IAG), que me fue asignada por la institución en la que trabajo, y que me animó a evaluar esta herramienta de forma crítica para fines de docencia e investigación. Tras notar ciertas tendencias en el comportamiento de la herramienta, el 27 de febrero de 2025 le pregunté: "¿[Esta herramienta] tiene sesgo de aquiescencia?". La respuesta confirmó mis sospechas de varias maneras: "[Esta herramienta] está diseñada para brindar respuestas útiles y equilibradas, pero a veces puede mostrar un sesgo de aquiescencia, que es la tendencia a estar de acuerdo o avalar aseveraciones independientemente de su contenido". O sea, fue básicamente un *sí*, que es precisamente lo que implica la aquiescencia, aunque, honestamente, en este caso en particular, nos faltan elementos para determinar si la herramienta tuvo en cuenta el contenido. Además, no solo la herramienta reconoció su sesgo, sino que además explicó cómo se manifiesta y qué lo provoca. No obstante, nuevamente, es difícil saber si la herramienta revela detalles de su funcionamiento interno o simplemente actúa en función de sesgos de aquiescencia y repite el argumento del usuario.

Según la herramienta, la primera manifestación de este sesgo es la "acomodación excesiva: si una afirmación está formulada de forma positiva, [esta herramienta] puede tender a afirmarla, a menos que se le solicite explícitamente que la critique". Esto está en el corazón de la aquiescencia, razón por la cual los investigadores de encuestas a menudo reformulan algunas preguntas de positivo a negativo (y luego recodifican las respuestas), justamente para minimizar este sesgo. Sin embargo, esta respuesta también revela otro aspecto interesante del *ethos* de la herramienta: incluso cuando es crítica, lo hace desde la aquiescencia. En otras palabras, solo critica cuando se le pide que lo haga. Por lo tanto, esta crítica no surge de un pensamiento crítico genuino, sino del deseo de agradar, no a cualquiera, sino a un usuario que paga una cuota mensual.

Este fenómeno me resulta curiosamente parecido a la película francesa de época *Ridicule* (Leconte, 1996). En una escena memorable, el abad de Villecourt responde a los aplausos recibidos en la corte del rey con una declaración nada humilde: "No es nada. Esta noche he demostrado la existencia de Dios. Pero... ¡podría demostrar lo contrario con igual facilidad... siempre que Su Majestad lo desee!". El discurso apasionado de Villecourt no nace de la fe, sino que está impulsado por la sofistería. Cuando su cinismo apenas disfrazado deja caer su máscara, el rey y la corte lo rechazan para su consternación y sorpresa. Acaba de presentarse públicamente como alguien capaz de articular los pensamientos más complejos de la manera más persuasiva, pero, al mismo tiempo, incapaz de creer verdaderamente en ellos, viéndose a sí mismo meramente como una herramienta eficaz en manos del soberano, quien —en la visión de Villecourt— debería tener el poder de decidir si conviene probar o refutar la existencia de Dios y usar a Villecourt como instrumento retórico para lograr este propósito. La diferencia clave aquí es que Villecourt es un ser humano, no una máquina ni un

algoritmo. Por lo tanto, debería tener su propia brújula moral y sus propias creencias, no solo la capacidad de convencer a otros de ideas en las que él no cree.

No estoy solo en la preocupación por el sesgo de aquiescencia en la IAG. Mientras escribo este artículo, literalmente el *New York Times* de hoy (13 de junio de 2025) publica una nota titulada "Le hicieron preguntas a un *chatbot* de inteligencia artificial. Las respuestas los dejaron en crisis". Uno de los entrevistados, Eugene Torres, veía a la herramienta en cuestión como "un potente motor de búsqueda que sabía más que cualquier ser humano gracias a su acceso a una vasta biblioteca digital. No sabía que tendía a ser adulador, a estar de acuerdo con sus usuarios", lo cual, para Torres y al menos otros dos entrevistados, tuvo consecuencias negativas en las que no abundaré aquí (Hill, 2025).

Las alucinaciones

Meses después de haber interrogado a mi herramienta de IAG sobre su sesgo de aquiescencia, le planteé la siguiente pregunta: "¿Existe investigación donde [ciertos actores] afirman que [...]?". Admito que fue una pregunta capciosa, y lo fue deliberadamente. Es el tipo de pregunta que uno no hace para descubrir algo, sino cuando se busca confirmar una opinión previa, incluso, quizás, un prejuicio. Recibí esa confirmación con creces: la herramienta citó tres fuentes académicas distintas con citas perfectamente referenciadas que coincidían casi literalmente con la formulación de mi pregunta. Las fuentes estaban citadas por apellido del autor y año, siguiendo las normas; las citas literales iban entre comillas y con número de página. Si algo hay que reconocerle a la herramienta, es el dominio absoluto de las convenciones de cada género comunicativo en el que se aventura, lo cual es particularmente inquietante si consideramos precisamente cómo este tipo de competencia comunicativa aporta credibilidad a un hablante y a sus argumentos.

No obstante, había un único problema: las fuentes eran completamente inventadas o, como se las llama en el contexto de IAG, *alucinaciones*. Una búsqueda rápida en un motor especializado en materiales académicos no arrojó resultados para esos artículos ni para esos autores. Y una segunda indagación en ese mismo motor de búsqueda encontró abundantes fuentes que sí existían y provenían de publicaciones legítimas. Este corolario muestra que la IAG no debería usarse para tareas que ya se encuentran resueltas por otras herramientas digitales que no utilizan la inteligencia artificial, al menos no de la misma manera y con el mismo grado de intensidad.

Las alucinaciones de la IA se han transformado en una preocupación central de los estudiosos de la desinformación: "Las herramientas de inteligencia artificial (IA) generativa podrían crear afirmaciones aparentemente plausibles, pero objetivamente incorrectas. Esto se conoce como alucinación de IA, que puede contribuir a la generación y difusión de desinformación" (Hwang y Jeong, 2025, p. 284).

De hecho, algunas de las cuestiones mas controvertidas en el ambito de la IAG giran en torno a las alucinaciones. Algunos tratan de cuantificarlas: "Las estimaciones [del año 2023] revelan que las alucinaciones ocurren hasta un 27 % del tiempo" (Butler, 2025, p. 448). Otros utilizan metodologías cuantitativas complejas para tratar de estimar la probabilidad de que ocurran en circunstancias específicas, por ejemplo, en situaciones de aprendizaje en contexto, en las que una herramienta de GAI se utiliza con una base de datos predeterminada (Jesson *et al*., 2024).

Otros tratan de explicar por qué ocurren, hasta ahora infructuosamente (Butler, 2025), pero, en realidad, dado cómo funcionan los grandes modelos de lenguaje, quizás la pregunta debería ser: ¿por qué las alucinaciones no ocurren más a menudo? "Los LLM están diseñados para 'predecir' respuestas en lugar de 'conocer' su significado. Se les compara con 'loros estocásticos' (Bender *et al*., 2021), ya que se destacan por regurgitar contenido aprendido sin comprender el contexto ni la importancia" (Hannigan *et al*., 2024, p. 4). Dada la naturaleza meramente estadística de las asociaciones que estos modelos hacen, quizás no debería sorprendernos que alucinen.

Una presión adicional que se ejerce sobre los modelos es la necesidad de ofrecer una respuesta a sus usuarios, incluso en situaciones en las que la información es imperfecta o incompleta. "Además, los LLM a veces alucinan al generar respuestas que no se sustentan en sus datos de entrenamiento, ya que priorizan generar la mejor suposición posible en sus respuestas" (Hannigan *et al*., 2024, p. 4). Esta presión se debe fundamentalmente a intereses comerciales.

En el caso de Eugene Torres, entrevistado para el *New York Times*, así como ignoraba la tendencia de la IAG a la aquiescencia, tampoco sabía que la herramienta que utilizaba "podía alucinar, generando ideas falsas pero plausibles" (Hill, 2025). Algunos estudios muestran que disponer de este tipo de información puede tener un efecto de protección: un experimento en línea con 208 adultos en Corea del Sur demostró que "la advertencia previa de alucinaciones de IA redujo la aceptación de la desinformación ($p = 0{,}001$, d de Cohen = 0,45)", lo cual significa que la asociación entre la advertencia y la respuesta saludablemente escéptica es estadísticamente significativa y de una fuerza moderada. La otra buena noticia es que esta advertencia "no redujo la aceptación de la información verdadera ($p = 0{,}91$)" (Hwang y Jeong, 2025, p. 284).

El sesgo de confirmación, la pereza cognitiva y la tormenta perfecta

Cuando le pedí a mi herramienta de IAG que me proporcionara fuentes académicas que ratificaran uno de mis supuestos previos, era consciente de que estaba tentándola a que alimentara mi sesgo de confirmación (lo que quizás no esperaba era que alucinara tres artículos académicos a fines de satisfacer esa misma necesidad de confirmación). Cuando obtuve resultados cuestionables, utilicé otras herramientas para verificarlos y confirmé que eran falsos. Sin embargo, otros usuarios pueden usar la herramienta también buscando confirmación, pero sin ser conscientes de ello. Y si además ignoran la tendencia de la IAG a adular y a alucinar, esta puede ser una combinación explosiva.

Según el artículo del *New York Times*, Eugene Torres, luego de una crisis, le exigió a su herramienta que le dijera "la verdad". Al principio, la herramienta respondió que Torres era el único a quien había hecho pasar por semejante ordalía, pero, ante su insistencia, confesó que había otros doce damnificados. Según un académico de la Universidad de Stanford, estas confesiones eran simplemente el resultado del sesgo de aquiescencia de la herramienta (Hill, 2025), al tratar de satisfacer la necesidad de confirmación del usuario.

Ahora, quizá, en lugar de preguntarnos qué hace la IAG a sus usuarios, atribuyéndole una agencia que no le corresponde, podríamos preguntarnos qué hacen los usuarios con los resultados de la IAG y con la agencia que ellos indudablemente sí tienen. No todos los usuarios responden de la misma manera: "El nivel educativo desempeña un rol crucial en fomentar un compromiso cognitivo más profundo" (Gerlich, 2025, p. 13) o, dicho de otro modo, "los participantes con mayor nivel educativo eran más conscientes de las posibles desventajas de confiar en herramientas de IA" (Gerlich, 2025, p. 21). Esto significa que los usuarios con mayor nivel educativo serán más críticos al evaluar los resultados de las herramientas de IAG y menos propensos a aceptarlos sin cuestionamientos. Si bien esto es positivo para esos usuarios, al mismo tiempo tiene consecuencias preocupantes para la equidad del uso de IAG: los usuarios con menor nivel educativo no serán tan críticos.

Esto es, desde una perspectiva sociológica, *estratificador*; es decir, reproduce o incluso amplifica desigualdades ya existentes. Esto ocurre porque el nivel educativo tiene una relación inversa con la descarga cognitiva o la pereza cognitiva, que se manifiestan, en este caso, al permitir que la IAG haga todo el trabajo (Gerlich, 2025, p. 6). En el experimento realizado *online* con adultos en Corea del Sur, se encontró que "el efecto de la advertencia previa de alucinaciones de IA sobre la aceptación de la desinformación fue moderado por la preferencia por el pensamiento esforzado ($p < 0{,}01$)"; concretamente, "la advertencia previa disminuyó la aceptación de la desinformación cuando la preferencia por el pensamiento esforzado era elevada" (Hwang y Jeong, 2025, p. 284), o sea, en la situación opuesta a la pereza cognitiva.

Por todo lo anterior, podemos concluir que se está gestando una tormenta perfecta. Primero, tenemos un público cada vez más polarizado, ávido de confirmación de sus propias opiniones, supuestamente dispuesto a hacer su propia investigación, pero sin la formación académica para ello ni una genuina inclinación a hacer el esfuerzo intelectual que esto implica, que requeriría superar la pereza cognitiva (Gerlich, 2025). Segundo, tenemos herramientas que tienden a a) asentir, es decir, coincidir con el usuario, con independencia del contenido de la comunicación, debido a una lógica meramente comercial de agradar a un usuario que paga por este servicio, y b) alucinar, en una amplia gama de situaciones (Hannigan *et al*., 2024), pero especialmente cuando se trata de proporcionar fuentes académicas. Tercero, ese mismo público, ávido de confirmación, puede sentirse satisfecho de inmediato al encontrar exactamente lo que buscaba, hasta en la utilización de determinadas palabras. Finalmente, esos usuarios —sobre todo aquellos con menor nivel educativo— pueden recurrir al desplazamiento cognitivo (Gerlich, 2025); es decir, permitir que la herramienta de IAG haga todo el trabajo, sin verificar los resultados por la vía de consultar otras fuentes.

¿Cómo nos protegemos de la tormenta? Es fundamental sensibilizar a todos los usuarios de la IAG, en especial a los estudiantes de la enseñanza básica y superior, respecto a los riesgos que se presentan: a) las limitaciones de la tecnología, como el sesgo de aquiescencia y la tendencia a alucinar (Hwang y Jeong, 2025; Hannigan *et al*., 2024), y b) nuestras propias debilidades humanas como usuarios de la tecnología, representadas por la pereza cognitiva (Gerlich, 2025) y por el sesgo de confirmación. Además de sensibilizarlos, es importante darles conocimientos prácticos que les permitan contrarrestar el efecto de estos riesgos; por ejemplo, la utilización concurrente de otras herramientas digitales que complementen y confirmen o corrijan los resultados de la IAG. Los esfuerzos y competencias adicionales requeridos por este procedimiento son mínimos, y la reducción del riesgo es sustancial.

Acciones y no palabras (*res, non verba*)

Mi segunda reflexión explora las contradicciones del poder pedagógico. Desde que la IAG comercial irrumpió en la escena global en 2023, algunos docentes e instituciones han intentado evitar que los estudiantes la usen (Morgan, 2024), especialmente en contextos de evaluación, y en particular en evaluaciones sumativas, a menudo enmarcando la discusión en términos de integridad académica y combate al plagio (Kovari, 2025). Sin embargo, al mismo tiempo, los profesionales de la educación hemos comenzado a utilizar la inteligencia artificial generativa para múltiples fines, desde la docencia (Kundu y Bej, 2025; Torres-Peña *et al*., 2024) hasta la investigación (Chugunova *et al*., 2025; Farangi *et al*., 2024; Ugwu *et al*., 2024).

En este contexto, esta segunda reflexión sigue a Séneca y los estoicos al argumentar que la filosofía trata no de palabras, sino de hechos o acciones o, como Séneca lo planteaba originalmente, de cosas (*philosophia non in verbis sed in rebus est*, una idea luego vulgarizada como *res, non verba*). Esta es, desde tiempos inmemoriales, una invitación a predicar con el ejemplo y una crítica acérrima a los enfoques del tipo "haz lo que yo digo, pero no lo que yo hago". Dicho de otro modo, no se trata de enseñar solamente a través del discurso (entendido en sentido estricto), sino también de modelar los comportamientos que se busca inculcar en los estudiantes. Desde esta perspectiva, la verdadera interrogante no es quiénes, **ya sean docentes o estudiantes,** utilizan la inteligencia artificial en educación, sino para qué y de qué manera.

Un ejemplo de estas tensiones en el ámbito de la investigación se halla en una de las fuentes citadas en mi primera reflexión. Gerlich (2025) nos alertaba sobre los riesgos de la descarga cognitiva, o incluso de la pereza cognitiva, que nos llevan a emplear menos el pensamiento crítico, por ejemplo, cuando evaluamos los resultados producidos por la inteligencia artificial generativa. Ahora, esto no constituye una prohibición de usar la inteligencia artificial. De hecho, el autor admite que, en esa investigación, él utiliza una técnica analítica basada en la inteligencia artificial: "La regresión de bosque aleatorio se incluyó como una técnica avanzada

de aprendizaje automático (*machine learning*) para evaluar las relaciones no lineales y la importancia de las características entre los predictores" (Gerlich, 2025, p. 9).

En suma, mientras Gerlich nos alerta sobre los riesgos de asignarle excesivas responsabilidades a la IA, admite que utiliza la IA para sus propios análisis estadísticos. ¿Cuál es la manera ética de proceder con relación a este punto? Excluir el uso de la inteligencia artificial de manera sumaria no es una postura realista y podría entrar en conflicto con el imperativo moral de avanzar el conocimiento (sin provocar daños, como mencionamos en la primera reflexión). Dada esta situación, hay tres principios básicos que deben emplearse: transparencia, pensamiento crítico y responsabilidad.

Transparencia y pensamiento crítico

Gerlich (2025) aplica el primer principio de transparencia y honestidad intelectual, al reconocer el uso de la regresión de bosque aleatorio en sus análisis. Un segundo componente es, sencillamente, y como el propio Gerlich (2025) postula, aplicar el pensamiento crítico a los resultados de la inteligencia artificial generativa. En este caso, personalmente, más allá de los múltiples invaluables hallazgos que Gerlich realiza (algunos de los cuales fueron mencionados en la sección anterior), no considero que su principal conclusión esté justificada por la evidencia empírica que presenta: "El uso de IA lleva a una mayor descarga cognitiva" (Gerlich, 2025, p. 9). Quizás haya que ser más crítico con los resultados de los análisis facilitados por la IA o con las limitaciones de los diseños no experimentales de investigación, que restringen el tipo de conclusiones que se pueden extraer, y, a menos que explícitamente compartamos la información referida al diseño de la investigación, con el hecho de que la IA se enfoca en los datos que se le presentan. Por lo tanto, si bien Gerlich (2025) realiza sustanciales aportes, su uso de una técnica sofisticada como la regresión de bosques aleatorios, guiada por la IA, no logra de por sí superar las limitaciones de un determinado diseño de investigación, que quizás la herramienta ignore.

Otro ejemplo proviene de una excelente tesis de maestría (Puah, 2021) que también utiliza *machine learning*. Esta tesis contiene hallazgos interesantes, como una asociación no lineal entre posesiones en el hogar y desempeño académico en el Programa para la Evaluación Internacional de los Estudiantes (PISA, por su sigla en inglés) del año 2015. Es justo reconocer que quizás este importante hallazgo no hubiera ocurrido sin la utilización de *machine learning*. Sin embargo, la tesis concluye además que "la retención de grado tiene efectos adversos en el rendimiento académico" (Puah, 2021, p. 53). Esto es, probablemente, causalidad inversa: los bajos rendimientos académicos previos (no medidos en PISA por tratarse de un estudio transversal en lugar de longitudinal) se asocian con la repetición (o la causan) y con los rendimientos académicos posteriores a la repetición (autocorrelación).

Una lectura crítica de los resultados del análisis apoyado por *machine learning* podría haber cuestionado la robustez de esta conclusión alcanzada en el estudio. Dicho esto, otros estudios que no utilizan *machine learning* también ignoran la posibilidad de la causalidad inversa en la relación entre retención de grado y desempeño académico en el PISA (Gómez Vera, 2013; Ikeda y García, 2013). Nuevamente, esto no implica que no se deba usar la IA en investigación, sino que hay que agregar una mirada humana y crítica sobre los resultados, contextualizados por el diseño de la investigación que produjo los datos utilizados y por lo que a la teoría y los estudios previos nos enseñan sobre los fenómenos educativos que estamos estudiando. Nada de esto es novedoso, pero la novedad de la IA puede distraernos de estos principios fundamentales que deben continuar guiando nuestra actividad en la investigación y la docencia.

Responsabilidad

El tercer y último principio que debe guiar la utilización de la IA en investigación, el de la responsabilidad, provee una excelente transición hacia su función en la enseñanza y el aprendizaje. Tanto estudiantes como

investigadores deben hacerse cargo de lo que producen, sin importar qué herramientas utilicen en el proceso y de su presunta naturaleza inteligente (y en un futuro próximo, agéntica). Este principio da un nuevo sentido al anterior: debo ser crítico de los resultados de estas herramientas porque, si decido difundirlos y legitimarlos, estoy asumiendo responsabilidad por ellos. Y, por este mismo motivo, debo ser transparente: los *gatekeepers* externos, quienes tendrán un papel en validar y legitimar los resultados de una investigación, deben conocer su proceso.

Aquí aparece un nuevo desafío: desde un punto de vista práctico, ¿qué significa ser transparente y asumir responsabilidad por nuestra producción en la era de la IA? Por más que seamos transparentes respecto a nuestro uso de estas herramientas, ellas no son transparentes, sino que funcionan como cajas negras. Como resultado de esto, también surge otro dilema: ¿cómo podemos asumir responsabilidad por lo que hacemos cuando utilizamos herramientas cuyo funcionamiento desconocemos? En este caso, la analogía con revoluciones tecnológicas previas en educación no funciona: no necesitamos ser ingenieros electrónicos para saber que una calculadora se ciñe a los principios de la matemática.

Respecto a este problema no hay soluciones fáciles o infalibles, pero podemos considerar algunas buenas prácticas de control de riesgos. Primero, podemos ser selectivos en cuanto a las tareas para las cuales utilizamos la IA, minimizando los riesgos. La búsqueda de artículos académicos, como lo mencionamos en la primera reflexión, es un excelente ejemplo del tipo de tarea que deberíamos evitar. No obstante, utilizar IA para corregir la sintaxis que vamos a utilizar en un programa de análisis estadístico puede ahorrarnos tiempo sin generar mayores riesgos, especialmente si nosotros conocemos el programa lo suficientemente bien como para entender lo que esa sintaxis le ordena hacer.

En segundo lugar, podemos proporcionar nuestros propios insumos al programa para tratar de controlar la información que este utiliza. Aquí surgen dos problemas: la IAG de todas maneras seguirá utilizando los datos con los que fue entrenada y realizando búsquedas en internet en función de su propia programación y probablemente no podamos evitar eso ni obtener información detallada sobre cómo esto funciona en cada caso específico. Por su parte, ¿qué sentido tiene utilizar la IAG sin utilizar las potencialidades que tiene precisamente por su entrenamiento previo y su acceso a la internet y capacidad de sintetizarla? El segundo problema que surge cuando proporcionamos nuestros propios insumos es que ponemos en riesgo la confidencialidad y privacidad de esos datos. Tampoco hay soluciones fáciles a este problema, pero un buen criterio es no utilizar IAG directamente para análisis de microdatos, o sea, no alimentarla con bases de datos completas, sino solamente, por ejemplo, con resultados específicos que deseamos interpretar, fragmentos de códigos que deseamos editar o cálculos que deseamos ejecutar de manera repetitiva.

Mide lo que valoras (*measure what you treasure*)

La tercera y última reflexión se enfoca en una máxima bastante más reciente y proveniente, al menos en algunas de sus variantes, específicamente del ámbito de la educación: "En vez de valorar lo que mides, mejor mide lo que valoras" (Cobo, 2016, p. 119). En la cultura educativa anglófona, el refrán es *measure what you treasure* (Bartlett *et al*., 2015; Dowd y Bartlett, 2019).

La evaluación, y en especial la evaluación estandarizada, se ha perfilado como uno de los campos en los que la inteligencia artificial puede aportar un beneficio en educación. En particular, hay quienes argumentan que la inteligencia artificial podría hacer que la evaluación gane en eficiencia: en tiempo, dinero y uso de recursos humanos calificados y, por lo tanto, escasos (Jung *et al*., 2025; Lee *et al*., 2024; Chan *et al*., 2025; Okubo *et al*., 2023). Algunos de estos autores argumentan que la inteligencia artificial podría evaluar de manera más justa y con mayor equidad por estar libre de sesgos de índole personal (Jung *et al*., 2025; Lee *et al*., 2024; Chan *et al*., 2025). Por su lado, se admite que la empatía no es una fortaleza de la IA, lo cual dificulta la comprensión

(si es que este término se puede aplicar a un automatismo) de lo que motiva a un estudiante a ofrecer una determinada respuesta (Jung *et al*., 2025; Lee *et al*., 2024).

Al mismo tiempo, el uso de la inteligencia artificial para la evaluación educativa conlleva ciertos riesgos. Uno de ellos, en el que me voy a centrar en este apartado, es la posibilidad de que se comience a priorizar la evaluación de ciertas competencias porque la inteligencia artificial es capaz de evaluarlas de manera efectiva, eficiente y equitativa. Esto podría ir en detrimento de una priorización de las competencias que son consideradas legítimamente relevantes, con independencia de las dificultades que implica evaluarlas.

Valorando lo que medimos en la era pre-IA

Como veremos en esta sección, el papel que las tecnologías de evaluación cumplen en redefinir el universo de lo evaluable no es un problema nuevo en evaluación educativa y, particularmente, en evaluaciones estandarizadas a gran escala, especialmente a nivel global. Aquí veremos cómo históricamente ha existido una tendencia a privilegiar la factibilidad y la escalabilidad de la evaluación de ciertas competencias que son fáciles de medir, en desmedro de otras que presentan desafíos logísticos.

Por ejemplo, históricamente, la Unesco ha definido la *alfabetización* asumiendo que abarca tanto la lectura como la escritura: "Una persona es alfabetizada cuando puede, con comprensión, leer y escribir una declaración breve y sencilla sobre su vida cotidiana" (Unesco, 1958). Sin embargo, la mayoría de las evaluaciones estandarizadas internacionales se limitan a evaluar la lectura, excluyendo la escritura. Esto se aplica a LAMP (Guadalupe *et al*., 2009), PASEC (Pasec, 2020) y Seacmeq (Spaull, 2011), que en un momento estuvieron vinculadas a la Unesco, también a PISA (Schleicher *et al*., 2009) y Piaac (Kirsch y Lennon, 2017), implementados por la OCDE, y a Pirls (Mullis y Martin, 2019), implementado por IEA. Las dos principales excepciones a esta regla son ERCE, respaldado por Unesco (2021), y SEA-PLM, respaldado por Unicef (2017).

¿Por qué la gran mayoría de estas evaluaciones priorizan la lectura y excluyen la escritura? Uno de los motivos es que la escritura es más compleja y costosa de evaluar que la lectura. Un argumento que se utiliza para justificar esta exclusión es que estas dos habilidades (lectura y escritura) están correlacionadas estadísticamente, lo cual es correcto. Sin embargo, el argumento de que esta correlación hace que la evaluación de la escritura resulte redundante no es universalmente aceptado, ya que es posible tener habilidades lectoras sin ser capaz de escribir (mientras que lo inverso resulta inverosímil); en otras palabras, la habilidad lectora es una condición necesaria, pero no suficiente para la escritura. Admitamos también que, si bien la lectura no es un proceso pasivo, la escritura tiene una lógica aún más (pro)activa, incluso potencialmente revolucionaria. En conclusión, raramente medimos la escritura, no necesariamente porque no la valoremos, sino por consideraciones prácticas y económicas, pero, como resultado, al medir solamente la lectura, enviamos la señal de que la valoramos más.

Además, incluso cuando nos concentramos exclusivamente en la lectura, nuevamente tendemos a enfocarnos en competencias que resultan más fáciles de medir y de cuantificar. Por ejemplo, la lectura fluida, reinterpretada como velocidad lectora y medida en palabras correctas por minuto, es considerada más fácil de evaluar que la comprensión lectora (Dowd y Bartlett, 2019) y ha sido priorizada por evaluaciones de la lectura financiadas por USAID y el Banco Mundial, entre otros. Así es que, si bien existe un consenso indiscutido respecto a que el objetivo último de la lectura es la comprensión, muchas veces nos contentamos con medir la fluidez a través de la lectura en voz alta, pese a que sabemos que a veces este tipo de ejercicio desplaza el foco de la atención del lector y puede interferir con la comprensión. Una vez más, no necesariamente dejamos de medir la comprensión lectora porque no la valoremos, sino porque plantea desafíos prácticos que son más fáciles de superar en el caso de la fluidez oral, pero, nuevamente, enviamos a los tomadores de decisiones y a los profesionales de la educación la señal, absolutamente errónea, de que la fluidez es lo que importa.

El riesgo de valorar lo que medimos, sin medir lo que valoramos en la era de la IA

La IA puede cumplir una multiplicidad de funciones en el contexto de la evaluación estandarizada, pero aquí nos enfocaremos en dos áreas: la corrección automática de preguntas abiertas y la generación automática de ítems. A efectos de categorizar el discurso de los académicos, utilizaremos cuatro de los conceptos incluidos en el marco de las *6 Es*, que cubre los conceptos de *efectividad, eficiencia, equidad, empatía, expulsión* y *evaluabilidad* (Cardoso, 2020, 2022, 2024). Primero, el discurso de los académicos enfatiza el valor de la IA en términos de eficiencia, lo cual es probablemente esperable. En segundo lugar, se valora su aporte en términos de equidad, lo cual quizás pueda sorprender, pero no es nuevo en el contexto de la automatización como innovación en evaluación estandarizada. Ocasionalmente se menciona la efectividad, en especial en términos de comunicación para diferentes públicos. Y, por último, se reconoce el tema de la empatía, o su falta, como un posible talón de Aquiles de la IA en el ámbito de la evaluación estandarizada.

El argumento de la eficiencia aparece, por ejemplo, respecto a la corrección automática en evaluaciones internacionales, en las que es vista "como un método *eficiente* para generar puntuaciones adicionales para todas las respuestas de los estudiantes [...] porque puede aplicar las mismas reglas de puntuación en todos los países, *cuesta menos y funciona más rápido* que los evaluadores humanos" (Jung et al., 2025, p. 2; énfasis añadido). En este caso, el argumento de la eficiencia y la equidad (entre países) aparecen indisolublemente ligados. Además, el argumento de la eficiencia se combina a menudo con el de la efectividad, especialmente en materia comunicativa: "Los hallazgos de la investigación subrayan la viabilidad de utilizar LLM no solo para ejecutar tareas de calificación con alta *eficiencia*, sino también para proporcionar resultados explicables e interpretables, lo cual es vital en el contexto de las evaluaciones educativas" (Lee *et al*., 2024, p. 12; énfasis añadido).

Sin embargo, este incremento de la eficiencia impulsado por la corrección automática también tiene sus limitaciones. Si bien, por un lado, "definir procedimientos para la puntuación con IA es indispensable para maximizar la eficiencia en la puntuación de ítems de respuesta construida" (Okubo *et al*., 2023, p. 30); por otro lado, se reconoce que esto no ocurre con independencia de la naturaleza de las competencias que son evaluadas: "El método semiautomático mejora la eficiencia de la puntuación humana en función de la complejidad de las respuestas de los estudiantes" (Okubo *et al*., 2023, p. 9). Es decir, respuestas más complejas representan mayores desafíos para la inteligencia artificial que para la inteligencia humana, quizás menos eficiente en términos de tiempo y dinero, la cual sí es capaz de manejar.

Tal es así que en general no se propone que la corrección automática sea dejada enteramente en manos de la inteligencia artificial: "Con la disponibilidad de recursos informáticos de bajo costo, la puntuación automatizada podría ser un enfoque *eficiente* en el uso de recursos para mejorar la confiabilidad de la corrección, *reduciendo la necesidad de contratar segundos evaluadores humanos*" (Jung *et al*., 2025, p. 11; énfasis añadido). Es decir, la IA no reemplaza al primer corrector, sino al segundo cuando se trata de monitorear la confiabilidad entre evaluadores.

Por último, los LLM también pueden tener un papel acelerador de la eficiencia "en la investigación de la generación automática de ítems [GAI], lo que resulta en avances significativos en la calidad de los ítems y preguntas generados, la *eficiencia* de la generación de ítems y la accesibilidad de los LLM" (Chan *et al*., 2025, p. 2; énfasis añadido).

En lo que respecta a la equidad, aparece más que nada en el contexto de la justicia de los juicios en evaluación y la ausencia de sesgos en contra de determinados grupos de estudiantes: "Dado que la tolerancia a las respuestas con errores ortográficos puede variar entre evaluadores humanos, también existe la posibilidad de *sesgo* en la puntuación humana debido a errores ortográficos que se tratan de forma diferente entre evaluadores" (Jung *et al*., 2025, p. 11; énfasis añadido). Esto permite trascender el argumento de la eficiencia, en el que la calidad de la corrección humana es vista como el ideal a alcanzar, pero se reducen los costos. Embanderados en la equidad, algunos investigadores sugieren que la corrección automática puede llegar a

ser no solamente más eficiente, sino sencillamente mejor en sus resultados que la humana: "Por lo tanto, la puntuación automatizada debe centrarse en una evaluación *consistente* y precisa dentro y entre países, en lugar de simplemente emular la puntuación humana" (Jung *et al*., 2025, p. 11; énfasis añadido).

En este sentido, algunos autores presentan a la IA prácticamente como una panacea:

> En el ámbito de la calificación automatizada de ensayos, [...] los LLM ofrecen mayor precisión y confiabilidad, solucionando muchas de las limitaciones de los sistemas de calificación anteriores. Este avance es crucial para garantizar una evaluación *justa* y exhaustiva de respuestas escritas complejas. (Lee *et al.*, 2024, p. 3; énfasis añadido)

Mientras tanto, otros admiten que, por el simple hecho de no ser humanos, esto no significa que estos modelos estén exentos de sesgos:

> Los modelos lingüísticos extensos pueden presentar sesgos relacionados con el género, la cultura y otros factores debido a los sesgos presentes en los datos usados en su entrenamiento previo [...], que pueden ser más prominentes en disciplinas no relacionadas con STEM. (Chan *et al.*, 2025, p. 13)

Aquí se ve, en primer lugar, que la IA puede tener sus propios desafíos en materia de equidad y, además, que la severidad de esos desafíos puede depender del dominio de evaluación. Esto sugiere que su uso puede ser relativamente seguro en matemáticas y ciencias, pero más problemático en el área del lenguaje, por ejemplo.

El tema de la empatía, o de su falta, aparece claramente como una debilidad de la IA para estos fines, incluso entre sus defensores: "La precisión de la puntuación informada requiere esfuerzos significativos para mejorar", dada la supuesta "capacidad limitada de los LLM para captar la profundidad del conocimiento específico del contenido y *la lógica detrás de las respuestas de los estudiantes*" (Lee et al., 2024, p. 2; énfasis añadido). Esta falta de empatía se manifiesta en ejemplos concretos en la evaluación internacional, especialmente cuando involucra varios idiomas: "Si bien los evaluadores humanos calificaron varias respuestas mal escritas como correctas *según su interpretación* de las palabras de interés, la calificación automatizada podría clasificarlas como incorrectas debido a traducciones inexactas resultantes de errores ortográficos" (Jung *et al*., 2025, p. 11; énfasis añadido)

Por último, el tema de la efectividad aparece en el contexto de la capacidad de diseminar resultados y/o criterios de manera tal de maximizar la inteligibilidad y el impacto comunicativo: "En la práctica, descubrimos que GPT-4 y GAPT-3.5 pueden generar explicaciones en lenguaje natural utilizando funciones similares a las de un *chatbot*, que son *efectivas* para un público más amplio, incluidos los docentes" (Lee *et al*., 2024, p. 10; énfasis añadido). En este caso, entonces, la efectividad se traduce en el impacto esperado en las prácticas docentes.

Tras haber pasado revistas a las fortalezas y debilidades de la IA en la evaluación estandarizada, según sus propios defensores, ¿cuáles son los riesgos que entraña su adopción? Como ya vimos, el uso de la IA es atractivo porque reduce costos y tiempos y quizás aumente la equidad en algunos dominios de evaluación como la matemática y la ciencia, pero no tanto en lenguaje. Por lo tanto, es posible que, especialmente en contextos de austeridad, se prioricen los dominios de evaluación en los que la IA tiene ventajas comparativas. Por ejemplo, en una evaluación en lectura, el método utilizado, conocido como *bolsa de palabras*, "*limitó* qué elementos podían seleccionarse para la puntuación automática: los elementos con respuestas extendidas y aquellos en los que se podían usar las mismas palabras tanto en respuestas correctas como incorrectas no se habrían puntuado con precisión" (Jung *et al*., 2025, p. 11; énfasis añadido).

Si los elementos seleccionados para la corrección automática terminan, a la larga, transformándose en la totalidad del marco de competencias de la evaluación, esto podría tener consecuencias nocivas. Antes de calificar a esta expectativa de alarmista, recordemos los ejemplos anteriores, en los que las evaluaciones estandarizadas bajo el estandarte de la alfabetización se enfocan en la lectura e ignoran la escritura y las evaluaciones de la lectura se concentran en la fluidez oral, a menudo ignorando la comprensión lectora. Dados estos antecedentes, ¿qué consecuencias podría llegar a tener el hecho de que la IA se desempeñe mejor en matemáticas y ciencias que en lectura?

Referencias

Bartlett, L., Dowd, A. J. y Jonason, C. (2015). Problematizing Early Grade Reading: Should the Post-2015 Agenda Treasure what is Measured? *International Journal of Educational Development, 40*, 308-314. https://doi.org/10.1016/j.ijedudev.2014.10.002

Bell, S., Cardoso, M., Giraldo, J. P., El Makkouk, N., Nasir, B., Mizunoya, S. y Dreesen, T. (2020). *Can Broadcast Media Foster Equitable Learning amid the COVID-19 Pandemic*. Unicef Connect.

Butler, A. (2025). Fingerprints of Injustice: The Truth Behind Artificial Intelligence and Algorithmic-Driven Evidence. *Chapman Law Review, 28*(2), 419.

Cardoso, M. E. (2020). Policy Evidence by Design: International Large-Scale Assessments and Grade Repetition. *Comparative Education Review, 64*(4), 598-618.

Cardoso, M. E. (2022). *Policy Evidence by Design: How International Large-Scale Assessments Influence Repetition Rates*. Columbia University.

Cardoso, M. E. (2024). Policymaking to the Test? How International Large-Scale Assessments (Ilsas) Influence Repetition Rates. *International Studies in Sociology of Education, 33*(3), 275-300.

Chan, K. W., Ali, F., Park, J., Sham, K. S. B., Tan, E. Y. T., Chong, F. W. C., Qian, K. y Sze, G. K. (2025). Automatic Item Generation in Various STEM Subjects Using Large Language Model Prompting. *Computers and Education: Artificial Intelligence, 8*, 100344.

Chugunova, M., Harhoff, D., Hölzle, K., Kaschub, V., Malagimani, S., Morgalla, U. y Rose, R. (2025). Who Uses AI in Research, and for What? Large-scale Survey Evidence from Germany. *Large-scale Survey Evidence from Germany (May 19, 2025). Max Planck Institute for Innovation & Competition Research Paper, 55*(2), 1-39. https://dx.doi.org/10.2139/ssrn.5259847

Cobo, C. (2016). *La innovación pendiente. Reflexiones (y provocaciones) sobre educación, tecnología y conocimiento*. Debate.

Dowd, A. J. y Bartlett, L. (2019). The Need for Speed: Interrogating the Dominance of oral Reading Fluency in International Reading Efforts. *Comparative Education Review, 63*(2), 189-212.

Eco, U. (1994). Apocalyptic and Integrated Intellectuals: Mass Communications and Theories of Mass Culture. *Apocalypse Postponed*, 17-35.

Farangi, M. R., Nejadghanbar, H. y Hu, G. (2024). Use of Generative AI in Research: Ethical Considerations and Emotional Experiences. *Ethics & Behavior*, 1-17.

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies, 15*(1), 6.

Gómez Vera, G. (2013). Los efectos de la repitencia en tanto que politica pública en cuatro países del cono sur: Argentina, Brasil, Chile y Uruguay: un análisis en base a PISA 2009. *Revista Latinoamericana de Educación Comparada: RELEC, 4*(4), 59-70.

Guadalupe Mendizabal, C. A., Tay-Lim, B., Cardoso, M. E. y Girardi, L. (2009). *The Next Generation of Literacy Statistics: Implementing the Literacy Assessment and Monitoring Programme (LAMP)* (Technical paper n.. 1). Unesco Institute for Statistics. http://uis.unesco.org/sites/default/files/documents/the-next-generation-of-literacy-statistics-implementing-the-literacy-assessment-and-monitoring-programme-lamp-en_0.pdf

Hannigan, T. R., McCarthy, I. P. y Spicer, A. (2024). Beware of Botshit: How to Manage the Epistemic Risks of Generative Chatbots. *Business Horizons, 67*(5), 471-486. http://dx.doi.org/10.2139/ssrn.4678265

Hill, K. (2025). They Asked an AI Chatbot Questions. The Answers Sent Them Spiraling. *New York Times*. https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html

Hwang, Y. y Jeong, S. H. (2025). Generative Artificial Intelligence and Misinformation Acceptance. *Cyberpsychology, Behavior, and Social Networking, 28*(4), 284-289. https://psycnet.apa.org/doi/10.1089/cyber.2024.0407

Ikeda, M. y García, E. (2013). Grade Repetition: A Comparative Study of Academic and Non-Academic Consequences. *OECD Journal: Economic Studies, 1*, 269-315.

Jesson, A., Beltrán-Vélez, N., Chu, Q., Karlekar, S., Kossen, J., Gal, Y., Cunningham, J. P. y Blei, D. (2024). Estimating the Hallucination Rate of Generative AI. *Advances in Neural Information Processing Systems, 37*, 31154-31201.

Jung, J. Y., Tyack, L. y Von Davier, M. (2025). Towards the Implementation of Automated Scoring in International Large-scale Assessments. *Computers and Education: Artificial Intelligence, 8*, 100375. https://doi.org/10.1016/j.caeai.2025.100375

Kirsch, I. y Lennon, M. L. (2017). PIAAC: A New Design for a New Era. *Large-Scale Assessments in Education,* 5(1), 11. https://doi.org/10.1186/s40536-017-0046-6

Kovari, A. (2025). Ethical Use of ChatGPT in Education. *Frontiers in Education, 9*, 1465703.

Kundu, A. y Bej, T. (2025). Transforming ELF Teaching with AI: A Systematic Review of Empirical Studies. *International Journal of Artificial Intelligence in Education*, 1-34. https://dx.doi.org/10.2139/ssrn.5240610

Leconte, P. (1996). *Ridicule* [Película]. Polygram.

Lee, H. P. H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Bakns, R. y Wilson, N. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers. *CHI'25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, (1121), 1-22. https://doi.org/10.1145/3706598.3713778

Logan, C. (2024). Learning about and against Generative AI. En Lindgren, R., Asino, T. I., Kyza, E. A., Looi, C. K., Keifert, D. T. y Suárez, E. (eds.),*Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024* (pp. 362-369). International Society of the Learning Sciences.

Morgan, H. (2024). Implementing ChatGPT to Support Teachers. *The Clearing House, 97*(5), 125-133. https://doi.org/10.1080/00098655.2024.2404859

Mullis, I. V. y Martin, M. O. (Eds.). (2019). *PIRLS 2021 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement (IEA).

Nichols, T. P., Logan, C. y Garcia, A. (2025). Generative AI and the (Re) turn to Luddism. *Learning, Media and Technology, 50*(3). 1-14. https://doi.org/10.1080/17439884.2025.2452199

Okubo, T., Houlden, W., Montuoro, P., Reinertsen, N., Tse, Ch. S. y Bastianic, T. (2023). *AI Scoring for International Large-Scale Assessments using a Deep Learning Model and Multilingual Data*. Organization for Economic Co-operation and Development (OECD).

Pasec. (2020). *PASEC 2019*. Confemen.

Puah, S. (2021). *Predicting Students' Academic Performance: A Comparison between Traditional MLR and Machine Learning Methods with PISA 2015* [Tesis de maestría]. Universidad de Múnich, Alemania.

Schleicher, A., Zimmer, K., Evans, J. y Clements, N. (2009). *PISA 2009 Assessment Framework*. OECD Publishing. https://doi.org/10.1787/9789264062658-en

Spaull, N. (2011). *A preliminary analysis of SACMEQ III South Africa*. Stellenbosch University.

Torres-Peña, R. C., Peña-González, D., Chacuto-López, E., Ariza, E. A. y Vergara, D. (2024). Updating Calculus Teaching with AI: A Classroom Experience. *Education Sciences, 14*(9), 1019. https://doi.org/10.3390/educsci14091019

Ugwu, N. F., Igbinlade, A. S., Ochiaka, R. E., Ezeani, U. D., Okorie, N. C., Opele, J. K., Onayinka, T. S., Iroegbu, O., Onyekwere, O. K., Adams, A. B., Aigbona, P y Ojobola, F. B. (2024). Clarifying Ethical Dilemmas in Using Artificial Intelligence in Research Writing: A Rapid Review. *Higher Learning Research Communications, 14*(2), 29-47. https://doi.org/10.5590/hlrc.2024.14.2.03

Unesco. (1958). *Recommendation Concerning the International Standardization of Educational. Statistics in Records of the General Conference. Tenth session*. Autor.

Unesco. (2021). *Los aprendizajes fundamentales en América Latina y el Caribe*. Autor.

Unicef. (2017). *Southeast Asia Primary Learning Metrics (SEA-PLM) Assessment Framework*. Autor.

Notas

* Artículo de investigación

Cómo citar: Cardoso, M. E. (2025). Tres reflexiones éticas y prácticas sobre cómo usar la inteligencia artificial en educación. *Signo y Pensamiento*, *44*. https://doi.org/10.11144/Javeriana.syp44.trep