Editorial

La confianza también se cita. La alucinación bibliográfica y los desafíos éticos de la inteligencia artificial generativa

Trust is also a citation. Bibliographic hallucination and the ethical challenges of generative artificial intelligence

Oswaldo Bolo-Varela
Universidad Nacional Mayor de San Marcos. Lima, Perú
oswaldo.bolo@unmsm.edu.pe
ORCID: 0000-0001-7335-043X

Raúl Castro-Pérez*
Universidad Científica del Sur. Lima, Perú
ccastroperez@cientifica.edu.pe
ORCID: 0000-0002-8986-0888

Citar como: Bolo-Varela, O. y Castro-Pérez, R. (2026). La confianza también se cita. La alucinación bibliográfica y los desafíos éticos de la inteligencia artificial generativa. *Desde el Sur*, *18*(2), 0028.

Uno de los rasgos más conflictivos en la producción contemporánea del conocimiento científico es el uso y abuso de la inteligencia artificial generativa (IA-G). Su uso está transformando el ecosistema del conocimiento y obligando a la comunidad académica a replantear sus fundamentos (Llorens-Largo y Molina-Carmona, 2026). Su abuso está evidenciando que, al delegar tareas a diversos modelos de lenguaje de gran escala (como ChatGPT, Gemini o Copilot), también se facilita la delegación de comportamientos poco éticos (Köbis *et al*., 2025). En el contexto de la producción y difusión académica, uno de estos comportamientos cuestionados es la invención de referencias bibliográficas, es decir, la falsificación de citas.

Esta práctica, también conocida como «alucinación bibliográfica» (Watson, 2024), consiste en que la IA-G proporciona referencias que parecen formal y explícitamente impecables —con autores plausibles, títulos verosímiles, nombres de revistas reales y estructuras de DOI correctas—, pero que, tras una comprobación rápida, efectiva y humana, no corresponden a ninguna publicación existente. No es un error menor: diversos

* Autor corresponsal: Raúl Castro-Pérez, Universidad Científica del Sur. Lima, Perú. Correo: ccastroperez@cientifica.edu.pe

estudios confirman la magnitud del problema (Bhattacharyya *et al*., 2023; Topaz *et al*., 2026; Zhao *et al*., 2026). Tampoco es un error solo atribuible al funcionamiento estadístico-predictivo con que operan las IA-G. De acuerdo con el Committee on Publication Ethics (COPE), uno de los organismos reguladores de la ética editorial a nivel internacional, la responsabilidad sobre el contenido de los manuscritos siempre recae íntegramente en los autores humanos, incluso en aquellas secciones elaboradas con asistencia de inteligencia artificial (COPE Council, 2024). Por esta razón, la elaboración de referencias mediante IA-G no representa una categoría ética nueva, sino una forma más de fabricación de datos, una conducta ya identificada como mala práctica y ampliamente sancionada en la comunidad académica internacional.

Las consecuencias de este fenómeno para el ámbito académico editorial son indudablemente más graves que los efectos previsibles para casos individuales. Entre los daños fácilmente identificables para el ecosistema editorial destacan las obvias reputaciones dañadas, la pérdida de las horas/hombre invertidas en el proceso editorial o el debilitamiento de los procesos de revisión por pares. Si profundizamos en las ramificaciones del problema, nos encontraremos también con la contaminación de bases de datos e índices, el riesgo para decisiones institucionales y políticas públicas y, por supuesto, la distorsión del conocimiento científico. Una implicancia decisiva, pero no tan discutida con eficacia y detalle, es el daño a nivel epistemológico que genera este tipo de prácticas. En la epistemología social del conocimiento científico, la construcción de la ciencia colaborativa contemporánea necesita de relaciones de confianza, es decir, requerimos del vínculo entre agentes a quienes se les puede exigir rendir cuentas por sus acciones (Koskinen, 2024). El uso opaco de la IA-G deteriora el establecimiento de dicha confianza entre los agentes y la producción del conocimiento. En otras palabras, la dependencia de herramientas que no pueden ser sometidas a los criterios convencionales de confianza científica erosiona las condiciones bajo las cuales la comunidad académica valida colectivamente el conocimiento.

Si bien el proceso de revisión por pares fue diseñado para evaluar la solidez argumentativa y metodológica de un trabajo, no lo fue necesariamente para auditar la existencia exacta de cada referencia listada, y esta

limitación ha quedado expuesta ante el volumen creciente de manuscritos con bibliografía contaminada. Por su parte, las herramientas automáticas de detección de texto generado por IA-G presentan tasas de error que las hacen inadecuadas como único mecanismo de control, pues su sensibilidad varía según el modelo generador, el idioma del manuscrito y el campo disciplinar, y pueden ser eludidas mediante reescritura o parafraseo (Linardon *et al*., 2025). Ante estas dificultades, será entonces responsabilidad de las revistas que los artículos con referencias alucinadas sean retractados y retirados; y de los autores, ser responsables del contenido íntegro del artículo, incluidas las referencias (Bauchner y Rivara, 2026).

En *Desde el Sur* hemos afrontado esta situación recientemente. El artículo *Alfabetización digital y ciudadanía resiliente: evidencia empírica de futuros docentes chilenos frente a la posverdad* de María Teresa Castañeda Díaz, Susana Riquelme Parra y Manuel Pereira Barahona, publicado en el volumen 17, número 4, de 2025, ha sido retractado debido a la identificación de referencias bibliográficas inexistentes. Luego de la investigación editorial respectiva, los autores del artículo declararon el uso inadecuado de herramientas de inteligencia artificial generativa para la construcción preliminar del aparato bibliográfico. A pesar del reconocimiento de los autores de las fallas en el proceso de elaboración y verificación del manuscrito, para el comité editorial de *Desde el Sur* estas explicaciones no permiten restablecer la confianza en la integridad del artículo ni ofrecer garantías suficientes sobre la confiabilidad del conjunto del trabajo. Por esta razón, siguiendo las recomendaciones del Committee on Publication Ethics (COPE) y las propias Políticas Éticas de nuestra revista, el comité editorial decidió retractar el artículo en mención [DOI: 10.21142/DES-1802-2026-0053].

Este caso ha llevado a que reforcemos nuestras políticas editoriales. Hemos actualizado nuestras normas éticas, disponibles en la web de la revista, detallando el uso correcto y permitido de la inteligencia artificial generativa. También, a partir de este número y en adelante, *Desde el Sur* exige a todos los autores una declaración de uso de IA-G. Esta exigencia es parte de la declaración jurada que los autores deben enviar cuando inician el proceso editorial, esto es, cuando envían su manuscrito para ser sometido a la evaluación por pares ciegos. En este documento, el autor

acepta que, en caso de haber usado IA-G, esta se limitó exclusivamente a mejorar la legibilidad y el lenguaje de la obra, incluyendo apoyo en la traducción, y en ningún caso se empleó como fuente directa de información. Asimismo, reconoce que el uso de IA-G para generar contenido sin verificación humana está prohibido y que esta no puede ser considerada autora del manuscrito. Finalmente, acepta que el incumplimiento de estas disposiciones puede afectar la aceptación del manuscrito y que, de haber usado la IA-G, su uso ha sido informado siguiendo el modelo de las normas editoriales de la revista.

En un contexto en el que la credibilidad de la ciencia está cada vez más en cuestionamiento, se necesitan esfuerzos renovados para fortalecer la integridad de la investigación académica. Confiamos en que *Desde el Sur* seguirá contribuyendo a esta tarea mediante políticas editoriales claras, procesos rigurosos y decisiones transparentes, incluso cuando estas resulten difíciles. La retractación de un artículo no es precisamente un éxito editorial; sin embargo, sí es el reconocimiento de que corregir públicamente los errores forma parte del compromiso con la ciencia. En tiempos de inteligencia artificial generativa, preservar la credibilidad en el conocimiento exige recordar que la ciencia es un proyecto colectivo sustentado en la confianza, la cual solo puede preservarse mediante la transparencia, la rendición de cuentas y la responsabilidad de quienes producen y comunican conocimiento.

Contribución de autoría

Oswaldo Bolo-Varela y Raúl Castro-Pérez cumplieron con todas las fases CRediT.

Fuente de financiamiento

Autofinanciado.

Potenciales conflictos de interés

Ninguno.

Declaración de uso de IA

No se usó ningún tipo de inteligencia artificial en la elaboración de este manuscrito.

REFERENCIAS BIBLIOGRÁFICAS

Bauchner, H. y Rivara, F. P. (2026). Fabricated references: a new threat to editorial integrity. *The Lancet*, *407*(10541), 1765-1766. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(26)00798-1/abstract

Bhattacharyya, M., Miller, V. M., Bhattacharyya, D. y Miller, L. E. (2023). High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus*, *15*(5). https://www.cureus.com/articles/158289-high-rates-of-fabricated-and-inaccurate-references-in-chatgpt-generated-medical-content#!/

COPE Council. (2024). COPE position. Authorship and AI. https://doi.org/10.24318/cCVRZBms

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F. y Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, *646*, 126-134. https://www.nature.com/articles/s41586-025-09505-x

Koskinen, I. (2023). We Have No Satisfactory Social Epistemology of AI-Based Science. *Social Epistemology*, *38*(4), 458-475. https://doi.org/10.1080/02691728.2023.2286253

Linardon, J., Jarman, H. K., McClure, Z., Anderson, C., Liu, C., Messer, M. (2025). Influence of Topic Familiarity and Prompt Specificity on Citation Fabrication in Mental Health Research Using Large Language Models: Experimental Study. *JMIR Mental Health*, *12*, e80371. https://mental.jmir.org/2025/1/e80371

Llorens-Largo, F. y Molina-Carmona, R. (2026). La verdad y lo veraz: La universidad ante la encrucijada de la IA generativa. *Revista Iberoamericana de Educación a Distancia (RIED)*, *29*(2), 63-80. https://revistas.uned.es/index.php/ried/article/view/47174

Topaz, M., Roguin, N., Gupta, P., Zhang, Z. y Peltonen, L.-M. (2026). Fabricated citations: an audit across 2·5 million biomedical papers. *The Lancet*, *407*, 1779-1781. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(26)00603-3/fulltext

Watson, A. P. (2024). Hallucinated Citation Analysis: Delving into Student-Submitted AI-Generated Sources at the University of Mississippi. *The Serials Librarian*, *85*(5-6), 172-180. https://doi.org/10.1080/0361526X.2024.2433640

Zhao, Z., Wang, Y., Stuart, T., De Vaan, M., Ginsparg, P. y Yin, Y. (2026). LLM hallucinations in the wild: Large-scale evidence from non-existent citations. https://arxiv.org/pdf/2605.07723

Oswaldo Bolo-Varela es profesor auxiliar de la Universidad Nacional Mayor de San Marcos e investigador especializado en memoria cultural, narrativas mediáticas y escritura contemporánea de no ficción en América Latina. Es autor de diversos artículos y capítulos de libro sobre memoria histórica, negacionismo y el fenómeno del terruqueo. Se desempeña como editor general de la revista Letras y editor asociado de Desde el Sur, desde donde impulsa iniciativas editoriales en estudios culturales, memoria y política.

Raúl Castro-Pérez es máster en Communication, Culture and Society por Goldsmiths University of London (2007), y actual doctorando en Comunicación por la Universidad Autónoma de Barcelona. Es decano de Comunicación en la Universidad Científica del Sur. Es director de Desde el Sur.

Recepción: 25/6/2026
Aceptación: 29/6/2026