Ciencia y Educación
(L-ISSN: 2790-8402 E-ISSN: 2707-3378)
Vol. 7 No. 12.1
Edición Especial XII 2026

HUMAN-IN-THE-LOOP EN CIENCIA AUMENTADA POR IA: SESGOS, VULNERABILIDADES Y EVALUACIÓN. REVISIÓN DE ALCANCE

HUMAN-IN-THE-LOOP IN AI-AUGMENTED SCIENCE: BIASES, VULNERABILITIES AND ASSESSMENT. A SCOPING REVIEW

Autores: [1]Andrés Guillermo Chérrez Ramírez.
[1]ORCID ID: https://orcid.org/0000-0001-6485-5073
[1]E-mail de contacto: acherrezr@unemi.edu.ec
Afiliación: [1*]Universidad Estatal de Milagro, Facultad de Posgrados, Maestría en Inteligencia Artificial para la Educación, (Ecuador).
Artículo recibido: 14 de Octubre del 2026.
Artículo revisado: 16 de Noviembre del 2026.
Artículo aprobado: 16 de Noviembre del 2026.
[1]Ingeniero en Sistemas de Computación e Informática, graduado de la Escuela Politécnica Nacional, (Ecuador). Magíster en Gestión de Proyectos, graduado de la Escuela Politécnica del Ejército, (Ecuador). Maestrante de la Maestría en Inteligencia Artificial para la Educación, en la Universidad Estatal de Milagro, (Ecuador). Especialista de Investigación, en la Universidad Técnica de Ambato (Ecuador).

Resumen

La supervisión humana de la inteligencia artificial generativa en la comunicación científica se invoca con frecuencia, pero se desconoce cómo se configura y qué evidencia la respalda. Esta revisión de alcance se preguntó qué configuraciones de supervisión humana, sesgos y vulnerabilidades describe la literatura reciente sobre la coproducción entre humanos e inteligencia artificial en la comunicación científica, y qué implicaciones tienen para la evaluación de la investigación. Siguiendo PRISMA-ScR, se buscó en Scopus y Web of Science Core Collection (2020–2026; inglés y español) y se realizaron dos búsquedas dirigidas. De 385 registros de bases de datos y 18 de búsquedas dirigidas, se evaluaron 65 obras a texto completo y se cartografiaron 40 fuentes; el análisis combinó frecuencias absolutas y relativas con una síntesis narrativa por ejes, sin evaluar su calidad. El 95,0 % de las fuentes se publicó en 2025–2026 y predominaron la revisión por pares y las decisiones editoriales. Se describieron configuraciones *human-in-the-loop*, en las que la persona interviene en cada decisión, *human-on-the-loop*, en las que supervisa y puede intervenir, y sistemas sin intervención humana; *human-over-the-loop* solo aparece como propuesta conceptual. La evidencia aborda el anclaje, los sesgos de los modelos como evaluadores y las vulnerabilidades adversariales, con tasas muy altas de inyección indirecta de instrucciones en condiciones experimentales. La «distorsión métrica» se apoya en evidencia cualitativa. Se concluye que la presencia humana declarada no basta para acreditar una supervisión eficaz; el juicio humano previo, la separación entre instrucciones y datos y la evaluación responsable requieren validación.

Palabras clave: Comunicación científica, Evaluación de la investigación, Inteligencia artificial generativa, Supervisión humana, Revisión por pares.

Abstract

Human oversight of generative artificial intelligence in scholarly communication is frequently invoked, yet how it is configured and what evidence supports it remain unclear. This scoping review asked which human oversight configurations, biases, and vulnerabilities recent literature describes regarding human–artificial intelligence coproduction in scholarly communication, and what implications they have for research assessment. Following PRISMA-ScR, Scopus and Web of Science Core Collection were searched (2020–2026; English and Spanish), and two targeted searches were conducted. Of 385 database records and 18 records from targeted searches, 65 works were assessed in full text and 40 sources were charted; the analysis combined absolute and relative frequencies with a narrative synthesis by axis, without appraising source quality. Overall, 95.0% of the sources were published in 2025–2026, and peer review and editorial decisions predominated. Human-in-the-loop

configurations, in which a person intervenes in each decision, human-on-the-loop configurations, in which a person supervises and can intervene, and systems without human intervention were described; human-over-the-loop appears only as a conceptual proposal. The evidence addresses anchoring, biases of models acting as evaluators, and adversarial vulnerabilities, with very high indirect prompt injection rates under experimental conditions. "Metric distortion" rests on qualitative evidence. Declared human presence is not enough to establish effective oversight; prior human judgment, separation of instructions from data, and responsible research assessment require validation.

Keywords: Scientific communication, Research evaluation, Generative artificial intelligence, Human oversight, Peer review.

Sumário

A supervisão humana da inteligência artificial generativa na comunicação científica é frequentemente invocada, más não se sabe como é configurada nem que evidência a sustenta. Esta revisão de escopo perguntou que configurações de supervisão humana, vieses e vulnerabilidades a literatura recente descreve sobre a coprodução entre humanos e inteligência artificial na comunicação científica, e que implicações têm para a avaliação da pesquisa. Seguindo o PRISMA-ScR, realizaram-se buscas na Scopus e na Web of Science Core Collection (2020–2026; inglês e espanhol) e duas buscas direcionadas. Dos 385 registros de bases de dados e 18 de buscas direcionadas, 65 obras foram avaliadas em texto completo e 40 fontes foram mapeadas; a análise combinou frequências absolutas e relativas com uma síntese narrativa por eixos, sem avaliar a qualidade das fontes. Do total, 95,0 % das fontes foram publicadas em 2025–2026, com predomínio da revisão por pares e das decisões editoriais. Descreveram-se configurações *human-in-the-loop*, em que a pessoa intervém em cada decisão, *human-on-the-loop*, em que supervisiona e pode intervir, e sistemas sem intervenção humana; *human-over-the-loop* aparece apenas como proposta conceitual. A evidência aborda a ancoragem, os vieses dos modelos como avaliadores e as vulnerabilidades adversariais, com taxas muito altas de injeção indireta de instruções em condições experimentais. A «distorção métrica» apoia-se em evidência qualitativa. Conclui-se que a presença humana declarada não basta para comprovar uma supervisão eficaz; o julgamento humano prévio, a separação entre instruções e dados e a avaliação responsável requerem validação.

Palavras-chave: Comunicação científica, Avaliação da pesquisa, Inteligência artificial generativa, Supervisão humana, Revisão por pares.

Introducción

La comunicación científica atraviesa una transformación en la que el artículo, concebido tradicionalmente como una «unidad cerrada», se inserta en flujos iterativos de publicación, revisión, retroalimentación y versionado (Rogel, 2025). La incorporación de la inteligencia artificial generativa (IA Gen), y en particular de los modelos de lenguaje de gran tamaño (LLM), plantea desafíos profundos a las prácticas de autoría y de construcción de sentido en la investigación (Watson et al., 2025). Más allá de la corrección lingüística, estos sistemas se emplean para sintetizar literatura (Anthuvan et al., 2025); (Guerard et al., 2026), redactar manuscritos completos (Yuan et al., 2026) y generar o anticipar evaluaciones de revisión por pares (Idahl y Ahmadi, 2025); (Thelwall y Yaghi, 2025).

Esta transferencia de tareas intelectuales hacia sistemas artificiales genera tensiones éticas, cognitivas y metodológicas para la integridad científica. La primera es la alucinación bibliográfica: los LLM pueden producir datos o referencias inexistentes con apariencia verosímil (Daoudi, 2025); (Fettiplace et al., 2025). En la evaluación de PaperOrchestrator, un sistema de redacción automatizada de

extremo a extremo, el 68 % de las entradas BibTeX generadas presentaba algún grado de alucinación, el problema más frecuente y severo identificado por sus autores (Yuan et al., 2026). La segunda tensión concierne a la supervisión humana. Los marcos de humano en el bucle (*human-in-the-loop*, HITL) presuponen que un experto puede detectar y corregir los errores del sistema, pero esa supervisión está condicionada por el sesgo de automatización, la conformidad excesiva con una valoración errónea del sistema y por el anclaje (Agudo et al., 2024); (Lazaros et al., 2026).

El fenómeno está descrito desde hace más de dos décadas en la investigación sobre factores humanos: el sesgo de automatización lleva a omitir errores que el sistema no señala y a seguir recomendaciones erróneas (Skitka et al., 1999), y la complacencia reduce la vigilancia cuando la automatización parece fiable (Parasuraman y Manzey, 2010). En dos experimentos con juicios simulados de culpabilidad, Agudo et al. (2024) observaron que emitir el juicio propio antes de conocer una recomendación algorítmica errónea se asociaba con mayor precisión que emitirlo después, patrón que interpretaron como un posible efecto de anclaje.

En el ámbito académico, un piloto exploratorio con estudiantes de máster describió sobreconfianza en las primeras salidas del modelo y una «ilusión de credibilidad» inducida por el tono fluido del texto generado (Guerard et al., 2026). Si estos mecanismos operan también en la revisión por pares es una de las cuestiones que examina esta revisión. La tercera tensión afecta a la gobernanza y a los indicadores de evaluación. A partir de un estudio cualitativo con 18 editores de revistas de disciplinas empresariales, Baciu et al. (2026) proponen el concepto de «distorsión métrica» para describir el debilitamiento de la relación entre los productos académicos medibles y el trabajo intelectual que representan. Desde una perspectiva sistémica, Rogel (2025) caracteriza las fábricas de artículos como patologías de acoplamiento entre la presión institucional por publicar y las economías de escala de la publicación digital, y Sharma y Sarode (2026) denominan PSEUDOPRINT a una forma de pseudorrevisión por pares en la que las aportaciones generadas por IA pueden dominar mientras se presentan como evaluaciones humanas. Ante estas tensiones, distintos trabajos proponen marcos para integrar los LLM en el trabajo académico sin renunciar al juicio experto (Anthuvan et al., 2025); (Guerard et al., 2026).

La literatura sobre sistemas HITL de alto riesgo sostiene que la automatización completa resulta insuficiente y describe modelos híbridos de tipo «centauro», en los que el juicio humano y el sistema se complementan (Lazaros et al., 2026), con el investigador como «ancla epistémica» que da sentido, asegura la exactitud y salvaguarda los estándares éticos (Anthuvan et al., 2025). Las síntesis previas más próximas no cubren este vacío: la revisión sistemática de Lazaros et al. (2026) examina los sistemas HITL en general, en distintos dominios de alto riesgo, y la revisión narrativa de Anthuvan et al. (2025) se centra en la escritura académica y en un marco de colaboración para la producción de conocimiento.

No se identificó, sin embargo, un mapa de cómo se configura la supervisión humana en las fases de la comunicación científica, de qué sesgos se han estudiado y con qué evidencia, ni de qué salvaguardas se proponen o se han evaluado. El objetivo general de esta revisión de alcance es mapear la literatura científica reciente sobre la coproducción humano-IA en la comunicación científica. La pregunta de revisión, formulada

con el marco PCC, fue: ¿qué configuraciones de supervisión humana, qué sesgos y vulnerabilidades y qué salvaguardas describe la literatura reciente (2020–2026) sobre la coproducción humano-IA en la generación, la síntesis y la evaluación de textos científicos, y qué implicaciones señala para la evaluación de la investigación?

De forma específica, se propone: (1) clasificar las configuraciones de supervisión humano-IA, ubicación del bucle (loop), granularidad de la interacción y fase editorial de los sistemas y marcos descritos para generar, sintetizar y evaluar textos científicos; (2) sistematizar la evidencia disponible sobre sesgos cognitivos, como el de automatización y el anclaje, en la supervisión humana de contenidos generados o evaluados con IA; y (3) describir las implicaciones para la evaluación de la investigación, incluida la distorsión métrica, y las salvaguardas de gobernanza propuestas, entre ellas la extensión del Manifiesto de Leiden planteada por Thelwall (2026).

Materiales y Métodos

Se realizó una revisión de alcance, diseño adecuado para describir el volumen, la naturaleza y las características de la evidencia en un campo emergente (Arksey y O'Malley, 2005); (Munn et al., 2018), informada según PRISMA-ScR (Tricco et al., 2018), con un diagrama de flujo adaptado de PRISMA 2020 (Page et al., 2021). La revisión no se registró de forma prospectiva ni dispone de un protocolo publicado, aunque las guías vigentes recomiendan elaborarlo y registrarlo antes de la búsqueda (Peters et al., 2022). El enfoque es descriptivo-cartográfico y de orientación pragmática: se buscó describir la extensión y la naturaleza de la evidencia, no estimar efectos ni jerarquizar su solidez. Se adoptó una taxonomía publicada (véase «Extracción de datos») en lugar de categorías inductivas para que las fuentes fueran comparables con una tipología previa de la supervisión humana. El autor investiga la coproducción humano-IA en la comunicación científica y, en la Discusión, formula una propuesta propia de supervisión por capas. Para limitar el sesgo de confirmación, las propuestas del autor se presentan separadas de los hallazgos del corpus, cada afirmación se contrastó con el texto completo de la fuente citada y las discrepancias de selección las resolvió una persona ajena a la autoría. Los criterios se formularon con el marco PCC (Peters et al., 2020). Población: investigadores, revisores, editores y comités de evaluación académica. Concepto: coproducción humano-IA (HITL, HOTL, HIC, HATL), sesgos de automatización y anclaje, distorsión métrica e integridad ética.

Contexto: comunicación científica, publicación editorial y evaluación de la calidad asistidas por IA generativa o LLM. (1) artículos y preprints con marcos de colaboración, flujos de procesamiento (*pipelines*) multiagente o arquitecturas para textos científicos; (2) estudios experimentales o de simulación sobre la precisión del experto humano frente a errores de la IA; (3) trabajos teóricos y regulatorios sobre gobernanza, ética, rendición de cuentas e indicadores cienciométricos; (4) publicaciones de 2020–2026 en inglés o español.

(1) automatización industrial o médica ajena a la comunicación académica; (2) documentos sin validación metodológica, cartas editoriales sin datos o literatura gris. Después de la selección original se precisaron cuatro reglas de agrupación, que se declaran como desviación respecto de los criterios iniciales: (a) el marco PCC prevalece y los criterios de inclusión 1–3 se leyeron como tipologías no exhaustivas; diez fuentes solo los satisfacen con esta lectura; (b)

las fuentes que cumplen el criterio 2 fuera del contexto de la comunicación científica se clasificaron como evidencia transferible cuando examinan un mecanismo cognitivo aplicable a la revisión o la edición, y se analizan aparte del corpus nuclear; (c) las que solo aportan marco conceptual de otros dominios pasaron a literatura contextual, fuera del mapa; (d) la exclusión 2 se aplicó literalmente a editoriales, comentarios y declaraciones sin datos, y las perspectivas y los protocolos que cumplen el PCC quedaron como casos frontera.

Las reglas se aplicaron a las fuentes incluidas en la selección original; los registros excluidos por «ámbito no científico» (n = 32) y los no seleccionados en las búsquedas dirigidas (n = 14), que la regla (b) podría volver elegibles, no se reevaluaron. Por la regla (c), tres fuentes se clasificaron como literatura contextual: un comentario sobre la atribución de responsabilidad en la IA médica (Ranisch, 2024), un documento de trabajo sobre la supervisión humana en la administración pública (Constantino, 2022) y un análisis conceptual de la supervisión humana de sistemas de IA (Langer et al., 2025). Se citan en la Discusión y no forman parte de los denominadores.

La identificación en bases de datos se realizó el 1 de agosto de 2026 en Scopus (n = 265) y Web of Science Core Collection (n = 120), con una cadena que combinó términos de supervisión humana, comunicación científica e IA generativa en título, resumen y palabras clave. El período declarado fue 2020–2026 y los idiomas admitidos, inglés y español; la cadena se formuló solo en inglés. Las cadenas literales ejecutadas no se conservaron; el Anexo A informa de los elementos documentados de cada búsqueda (bases, fecha, campos, bloques conceptuales, límites y resultados). Se realizaron además dos búsquedas dirigidas (otros métodos) para cubrir mecanismos que la cadena principal no capturaba: B1, sobre normalización Unicode y seguridad de los sistemas editoriales, con sintaxis de Scopus limitada a 2025–2026 porque la inyección de instrucciones en sistemas editoriales se documenta a partir de 2025 (7 registros), y B2, sobre fatiga cognitiva en la supervisión (11 registros). Los registros se seleccionaron con los mismos criterios de elegibilidad que la búsqueda principal.

La selección se realizó en tres etapas: eliminación de duplicados, cribado de título y resumen frente al PCC y evaluación a texto completo (Page et al., 2021). En ambas fases, el autor decidió la elegibilidad de cada registro y Claude (Anthropic; modelo claude-opus-5) propuso decisiones de forma independiente; una persona ajena a la autoría resolvió las discrepancias. Las propuestas de la herramienta se trataron como sugerencias y no como una revisión humana independiente. No se calculó la concordancia entre decisiones: como las propuestas de la herramienta no constituyen una revisión humana independiente, un índice de concordancia entre el autor y la IA no equivaldría a la fiabilidad entre revisores que contempla PRISMA-ScR.

Esta omisión se declara como limitación. Se registraron identificación y tipo documental; diseño y tipo de evidencia; dominio y contexto; configuración del loop; granularidad; fase editorial; sesgos y salvaguardas con su respaldo empírico; y hallazgos y limitaciones declaradas. La taxonomía adapta la de Lazaros et al. (2026): ubicación del loop, granularidad y características temporales, con la dimensión temporal sustituida por la fase del ciclo editorial; las cinco configuraciones (*in-*, *on-*, *over-*, *under-* y *along-the-loop*) siguen la

tipología de Singh y Szajnfarber (2025, como se citó en Lazaros et al., 2026). Según esa tipología, *human-in-the-loop* designa configuraciones en las que la participación humana es necesaria para el funcionamiento del sistema; *human-on-the-loop*, aquellas en las que la supervisión humana es opcional pero permite intervenir; *human-over-the-loop*, aquellas en las que las personas controlan los objetivos y las restricciones del sistema; *human-under-the-loop*, aquellas en las que el sistema de IA controla o influye en la conducta humana; y *human-along-the-loop*, aquellas en las que personas y sistemas realizan tareas relacionadas en paralelo, con interacción lateral (Lazaros et al., 2026).

Se codificó la configuración descrita explícitamente o inferible del flujo, distinguiendo ambos casos (explícita, cuando la fuente usa el término o describe el punto de intervención humana; inferida, cuando se deduce del flujo de trabajo descrito), con codificación múltiple. Los documentos normativos, conceptuales, de encuesta o de consenso que no describen un sistema en funcionamiento se codificaron como «no determinable» en la configuración y como «políticas o gobernanza» en la granularidad; las revisiones con taxonomías que abarcan todas las configuraciones y las fuentes que no describen el flujo humano-IA se codificaron como «no determinable» en ambas dimensiones.

Las fases editoriales se asignaron con un vocabulario cerrado de seis categorías no excluyentes, en el que la asignación de revisores y editores forma parte de la revisión por pares y la decisión editorial. La extracción y la codificación se realizaron con asistencia de Claude (Anthropic); se revisaron las propuestas antes de incorporarlas a la matriz. Las características de cada fuente cartografiada (diseño, objeto de estudio, configuración, granularidad, fase editorial, sesgos y salvaguardas, y localizador en la fuente). Se calcularon frecuencias absolutas y relativas sobre las fuentes cartografiadas (n/N), conforme a las recomendaciones para el análisis y la presentación de resultados en revisiones de alcance (Pollock et al., 2023); con codificación múltiple cada fuente se cuenta una vez por categoría, por lo que los porcentajes pueden superar el 100 %. La síntesis narrativa se organizó en los tres ejes de los objetivos, y los resultados de la evidencia transferible se presentan separados de los del corpus nuclear. No se realizó evaluación crítica de la calidad, opcional en PRISMA-ScR (Tricco et al., 2018): los resultados describen la existencia y la naturaleza de la evidencia, no su solidez.

Resultados y Discusión

La búsqueda en bases de datos identificó 385 registros (Scopus, 265; Web of Science, 120). Tras eliminar 126 duplicados, se cribaron 259 registros por título y resumen y se excluyeron 106. De los 153 informes buscados a texto completo, cuatro no pudieron recuperarse. De los 149 evaluados, se excluyeron 87 por ausencia de metodología clara (n = 34), ámbito no científico (n = 32) o literatura gris (n = 21), y se seleccionaron 62.

Las búsquedas dirigidas complementarias sobre supervisión humana, sesgos, vulnerabilidades y evaluación de la investigación identificaron 18 registros adicionales. Aplicando los mismos criterios de elegibilidad y eliminación de duplicados, se seleccionaron cuatro por su pertinencia para la pregunta de revisión, independientemente del sentido de sus resultados. En conjunto, ambas vías reunieron 66 informes. El cotejo documental permitió identificar y eliminar una copia duplicada, por

lo que quedaron 65 obras únicas. En la evaluación detallada del texto completo, 22 obras no cumplieron los criterios de elegibilidad y 43 sí los cumplieron; de estas, 3 se clasificaron como literatura contextual (regla c) y 40 se cartografiaron, una de ellas como evidencia transferible (Figura 1). Por separado, se incorporaron 47 referencias de fundamentación metodológica y 2 de fundamentación conceptual sobre el sesgo de automatización. Estas y las 3 contextuales se excluyeron de los denominadores. La bibliografía comprende, por tanto, 52 referencias: 40 fuentes cartografiadas, 3 contextuales, 7 metodológicas y 2 conceptuales.

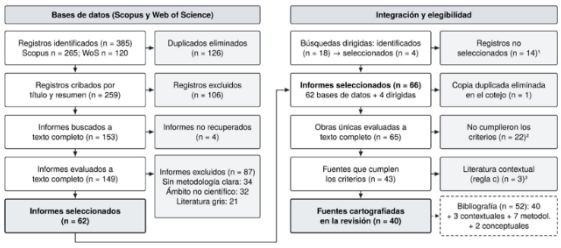


Figura 1. *Diagrama de flujo de la selección de fuentes (adaptado de PRISMA 2020).*

Fuente: Elaboracion propia.

Tres fuentes están publicadas en español (Marín y Martínez, 2025); (Pérez et al., 2026); (Rogel, 2025). Entre ellas, Marín y Martínez (2025) analizan 17 revistas colombianas, de las cuales dos (11,8 %) mencionan explícitamente la IA en sus normas editoriales. La Tabla 1 resume la distribución de las fuentes por configuración, granularidad y fase. La configuración se describe de forma explícita en una proporción limitada de las fuentes: de las 26 con una configuración determinable, 13 la describen explícitamente y 13 permiten inferirla del flujo de trabajo. Entre las 13 que describen supervisión humana, solo 6 lo hacen de forma explícita (4 *human-in-the-loop*, 1 *human-on-the-loop* y 1 *human-over-the-loop*). Las categorías *human-under-the-loop* y *human-along-the-loop* aparecen en una fuente de carácter taxonómico (2,5 %). De las 40 fuentes cartografiadas, 38 (95,0 %) se publicaron en 2025 o 2026; 36 son publicaciones en revistas, 2 corresponden a contribuciones en actas y 2 a preprints. Según su contribución principal, se distinguen 17 estudios empíricos (42,5 %), 14 desarrollos y evaluaciones técnicas de sistemas (35,0 %), 7 trabajos conceptuales o normativos (17,5 %) y 2 revisiones de literatura (5,0 %). Al clasificar la configuración funcional predominante descrita o inferida del flujo de trabajo, 7 fuentes corresponden a *human-on-the-loop* (17,5 %), 45 a *human-in-the-loop* (12,5 %) y 13 evalúan tareas automatizadas sin supervisión humana durante su ejecución (32,5 %).

Una fuente propone una arquitectura *human-over-the-loop* (2,5 %), de carácter conceptual y aplicada al diagnóstico educativo. En 14 fuentes no resulta determinable una configuración operativa predominante (35,0 %): 12 son documentos normativos, conceptuales, de encuesta o de consenso, una es una revisión taxonómica y en otra no se describe el flujo humano-IA. La comparación de resultados algorítmicos con evaluaciones humanas no se contabilizó, por sí sola, como supervisión durante la tarea.

Por objeto principal de estudio, la revisión por pares, la retroalimentación y las decisiones editoriales asistidas por IA reúnen 10 fuentes; la evaluación de la investigación, la calidad, la novedad y las métricas, 7; la ética, la gobernanza y las políticas científicas o editoriales, 7; la escritura científica, la autoría y la asistencia a la investigación, 6; la asignación automatizada de revisores y editores, 4; la supervisión humana, la agencia y la responsabilidad, 3; y la integridad de la revisión

y las vulnerabilidades adversariales, 3. Una de las fuentes aporta evidencia experimental transferible procedente de tareas de justicia penal simulada (Agudo et al., 2024). La revisión por pares y la decisión editorial constituyen la fase más representada (30/40; 75,0 %), seguidas de la redacción y autoría (12/40; 30,0 %), la publicación y sus políticas (11/40; 27,5 %), el cribado editorial (9/40; 22,5 %), la evaluación de la investigación (6/40; 15,0 %) y la preparación y gestión de la literatura (2/40; 5,0 %). Tres fuentes, la evidencia transferible y dos de alcance transversal, no corresponden a una fase editorial propia. Estas categorías no son excluyentes, pues una misma fuente puede abordar varias fases.

En cuanto a la granularidad predominante, 45 fuentes sitúan la intervención humana en hitos o puntos de control (12,5 %), 8 en la revisión o supervisión posterior de resultados (20,0 %) y 12 en políticas o mecanismos de gobernanza (30,0 %). Otras 13 evalúan tareas automatizadas sin intervención humana durante su ejecución (32,5 %), mientras que en 2 no puede determinarse la granularidad con la información disponible (5,0 %).

En el subconjunto de fuentes técnicas y empíricas directamente relacionadas con la comunicación científica (n = 30), se identifican 14 desarrollos o evaluaciones técnicas (46,7 %) y 16 estudios empíricos (53,3 %). La revisión por pares y la decisión editorial siguen siendo la fase más representada (24/30; 80,0 %). Predomina la automatización de la tarea evaluada (13/30; 43,3 %), seguida de *human-on-the-loop* (7/30; 23,3 %) y *human-in-the-loop* (3/30; 10,0 %); siete fuentes no permiten determinar una configuración operativa predominante (23,3 %). Catorce de las 40 fuentes (35,0 %) se clasificaron como desarrollos y evaluaciones técnicas de sistemas.

En la redacción, PaperOrchestrator descompone la escritura en siete etapas con un estado compartido y obtuvo una tasa de compilación LaTeX del 90,0 %, frente al 69,0 % ChatPaper (Yuan et al., 2026); opera sin puntos de control humano durante la generación, la evaluación humana es posterior, y sus autores documentan que el 68 % de las entradas BibTeX presentaba algún grado de alucinación, de modo que estos resultados no permiten afirmar que la arquitectura multiagente resuelva ese problema.

Research LLM organiza la investigación en tres etapas con puntos de control humanos explícitos y técnicas de verificación (Guerard et al., 2026), y el marco conceptual Scholarly HI-AI Loop, propuesto en un preprint, reparte siete etapas en las que la IA precriba y sintetiza mientras el investigador fija el alcance, refina y valida éticamente el texto (Anthuvan et al., 2025).

Tabla 1: *Distribución de las 40 fuentes cartografiadas por configuración, granularidad y fase editorial.*

Dimensión y categoría	n (%)
Configuración funcional predominante	
Tarea automatizada sin supervisión durante su ejecución	13 (32,5)
Human-on-the-loop	7 (17,5)
Human-in-the-loop	5 (12,5)
Human-over-the-loop (conceptual)	1 (2,5)
No determinable	14 (35,0)
Granularidad de la intervención humana	
Sin intervención durante la ejecución	13 (32,5)
Revisión o supervisión posterior de resultados	8 (20,0)
Políticas o mecanismos de gobernanza	12 (30,0)
Hitos o puntos de control	5 (12,5)
No determinable	2 (5,0)
Fase editorial (categorías no excluyentes)	
Revisión por pares y decisión editorial	30 (75,0)
Redacción y autoría	12 (30,0)
Publicación y políticas	11 (27,5)
Evaluación de la investigación	6 (15,0)
Cribado editorial	9 (22,5)
Preparación y gestión de la literatura	2 (5,0)

Fuente: Elaboración propia.

En la revisión por pares, los sistemas generan dictámenes (Idahl y Ahmadi, 2025), resumen artículos para los revisores (Majadly y Last,

2025), valoran la calidad de los informes (Sun et al., 2026) o estiman la novedad de las publicaciones (Kruzenshtern et al., 2026), predicen decisiones editoriales (Pandey et al., 2025); (Pendyala et al., 2025); (Thelwall y Yaghi, 2025) y asignan revisores o editores (Bagheri et al., 2026); (Bhattacharya et al., 2025); (Carpenter et al., 2025); (Xu et al., 2026). OpenReviewer, afinado con 79 000 revisiones expertas, presentó en 400 artículos de prueba una media y una dispersión próximas a las de las puntuaciones humanas (5,4 ± 1,1 frente a 5,4 ± 1,2), mientras que los LLM generalistas obtuvieron puntuaciones medias entre 6,9 y 8,1, en una escala de 1 a 10 (Idahl y Ahmadi, 2025).

En la evaluación de la investigación, las puntuaciones de calidad generadas por LLM se compararon con valoraciones expertas o indicadores derivados de ellas: en 104.187 artículos, las puntuaciones de un modelo de pesos abiertos mostraron correlaciones positivas, generalmente débiles, con las puntuaciones medias departamentales utilizadas como indicador indirecto de calidad individual, en las 34 unidades de evaluación (Thelwall, 2025).

En 2780 artículos de medicina, salud y ciencias de la vida, los modelos evaluados con más de 4000 millones de parámetros mostraron capacidad sustancial para estimar la calidad, especialmente al promediar varias puntuaciones (Thelwall y Mohammadi, 2026). Un *pipeline* de evaluación de la pertinencia semántica de las citas puntuó 5615 pares cita-referencia; en una submuestra estratificada de 150 pares, obtuvo una correlación de Spearman de 0,643 con la clasificación por mayoría de tres expertos (Isik y Guleryuzlu, 2026). En las 40 fuentes cartografiadas, 29 (72,5 %) abordan sesgos o riesgos de dependencia de la IA; 8 fuentes (20,0 %) aportan observaciones primarias sobre patrones de juicio o evaluación. Las categorías, no excluyentes, comprenden los sesgos sociales y de representación, incluidos los de idioma o región (19; 47,5 %); los sesgos y desviaciones de los LLM como evaluadores (13; 32,5 %, con observaciones empíricas en seis); la dependencia excesiva o complacencia (8; 20,0 %, estudiada empíricamente en dos); y los sesgos de revisores humanos o de los procesos institucionales de revisión y selección (10; 25,0 %).

El sesgo de automatización se aborda en 3 fuentes (7,5 %), el anclaje en 2 (5,0 %), ambas con observaciones empíricas, y la confirmación en 4 (10,0 %). No se identificaron mediciones específicas del sesgo cognitivo de conformidad. La evidencia experimental sobre el anclaje procede de la fuente transferible. En dos experimentos con juicios simulados de culpabilidad (N = 150, con participantes españoles, y N = 260, con participantes hispanohablantes de distintos países), Agudo et al. (2024) observaron que quienes emitían su juicio antes de recibir una recomendación algorítmica errónea eran más precisos que quienes lo emitían después.

Sin embargo, durante la fase con apoyo erróneo del segundo experimento, el porcentaje de participantes que acertó en los tres casos fue del 29,5 % en el grupo que recibió primero la recomendación y del 31,3 % en el que emitió primero su juicio, sin diferencias significativas (p = 0,765). Los autores interpretan los resultados como un posible efecto de anclaje y consideran aconsejable emitir un juicio explícito antes de mostrar la valoración del sistema, aunque esta secuencia no evitó el deterioro posterior de las decisiones. En el ámbito académico, el piloto cualitativo exploratorio de Guerard et al. (2026) describió

anclaje en las primeras salidas del modelo y la «ilusión de credibilidad». No se identificaron estudios que evaluaran experimentalmente el anclaje en revisores o editores. Las evaluaciones empíricas documentan sesgos y limitaciones de los LLM como evaluadores, que deben distinguirse entre sí: ChatGPT no predijo los resultados de revisión de F1000Research a partir de títulos y resúmenes (ρ de Spearman = 0,00), con correlaciones débiles en SciPost Physics y moderadas en ICLR (Thelwall y Yaghi, 2025); los LLM generalistas puntuaron por encima de los humanos (Idahl y Ahmadi, 2025).

Un LLM local mostró una concordancia leve y no significativa con la decisión editorial (κ = 0,126), numéricamente inferior a la de los revisores humanos agrupados (κ = 0,181), aunque el 90,5 % de sus recomendaciones conjuntas de rechazo o transferencia a una revista del mismo grupo editorial coincidió con alguna de esas dos decisiones finales (Patel et al., 2026); y, en una comparación de dos versiones de un manuscrito, los dos revisores humanos prefirieron la versión humana, mientras que el revisor de IA prefirió la generada por IA (Nathani et al., 2025). Estas diferencias de desempeño no demuestran, por sí solas, un mecanismo cognitivo específico.

La influencia de la IA en la revisión real también está documentada: se estimó que al menos el 15,8 % de las revisiones de ICLR 2024 se escribió con asistencia de IA y, en un análisis con emparejamiento, los envíos cercanos al umbral que recibieron una revisión asistida presentaron una probabilidad de aceptación 4,9 puntos porcentuales mayor (Russo et al., 2025). La coincidencia entre los aspectos señalados por GPT-4 y por los revisores humanos fue comparable a la observada entre dos revisores humanos (30,85 % frente a 28,58 % en revistas de la familia Nature) (Liang et al., 2024). En un ensayo aleatorizado con más de 20 000 revisiones de ICLR 2025, el 27 % de los revisores que recibió retroalimentación automatizada actualizó su revisión (Thakkar et al., 2026). La recepción y valoración de comentarios en el estudio de usuarios de Liang et al. (2024) corresponde a una configuración *human-on-the-loop*, mientras que la revisión iterativa de Thakkar et al. (2026) corresponde a una configuración *human-in-the-loop*. Ninguno de estos estudios de revisión académica comparó experimentalmente el orden de exposición entre el juicio humano y la sugerencia algorítmica.

Torrielli et al. (2026) evaluaron la inyección indirecta de instrucciones (*prompts*) con 100 manuscritos de OpenReview, cinco familias de cargas maliciosas (*payloads*) y 42 000 salidas de GPT-5.3 Instant y Gemini-3-flash-preview (21 000 por sistema): las tasas agregadas de éxito fueron del 98,34 % y del 94,02 %, respectivamente. Los autores denominan «ceguera contextual» a la incapacidad de los modelos para separar el contenido científico del texto de control incrustado.

En una comparación adicional con 26 manuscritos, su variante *Policy Puppetry*, en la ubicación de inyección con mejor desempeño, alcanzó tasas del 99,8 % y del 99,4 % en la categoría de detección mediante marcas, frente al 58,7–74,8 % de los *payloads* de Collu et al. (2026, como se citó en Torrielli et al., 2026), reevaluados en ese mismo estudio. Bajo instrucciones neutrales de revisión, el texto invisible elevó las recomendaciones de aceptación del 0 % al 99,2–100 % (Choi et al., 2025). Una marca de agua basada en una cita inventada alcanzó una inserción del 98,6 % mediante texto blanco; además, las pruebas de paráfrasis mostraron una persistencia de las

marcas superior al 94 % (Rao et al., 2025). Las recomendaciones incluyen sondas defensivas sujetas a políticas explícitas, controles que no dependan solo de detectores y saneamiento del contenido. No se identificaron evaluaciones del efecto de la conversión de archivos en plataformas editoriales reales. Baciu et al. (2026), mediante un estudio cualitativo exploratorio con 18 editores y editores asociados, describen fragmentación de políticas, ambigüedad sobre autoría y rendición de cuentas y desalineación entre las políticas editoriales y la evaluación institucional. Proponen el concepto de «distorsión métrica» para interpretar el debilitamiento de la relación entre los productos científicos cuantificables y el trabajo intelectual que se les atribuye, con las limitaciones de una muestra pequeña y basada en percepciones.

Thelwall (2026) considera que los diez principios del Manifiesto de Leiden siguen siendo pertinentes para la evaluación basada en citas, en juicio experto y en LLM entre ellos, que lo cuantitativo apoye la valoración cualitativa de expertos (Hicks et al., 2015, como se citó en Thelwall, 2026) y propone dos adicionales: análisis costo-beneficio de los métodos de evaluación y cumplimiento de los derechos de autor.

Rogel (2025) advierte que, con métricas calculadas sobre corpus contaminados por fábricas de artículos, plagio o textos generados por IA, el sistema pierde capacidad para distinguir comunicaciones válidas de simuladas. El análisis de las políticas de cinco editoriales identificó cuatro valores consistentes: autoría humana exclusiva, divulgación obligatoria, responsabilidad no delegable y confidencialidad (Le Duc, 2026). Un consenso Delphi acordó que la IA no puede figurar como autora y que los autores responden de todo el contenido asistido (Fettiplace et al., 2025), y Resnik y Hosseini (2025) sostienen que los LLM carecen de capacidades esenciales para la agencia moral. En los contextos editoriales estudiados, el uso descrito de IA se concentra en la evaluación preliminar y muchas revistas carecen de una declaración explícita sobre su utilización (Pérez et al., 2026); en la síntesis de Anthuvan et al. (2025), 12 de los 38 estudios examinados (31,6 %) no divulgaban su uso. Sharma y Sarode (2026) advierten del riesgo de PSEUDOPRINT, entendido como una revisión aparentemente humana que encubre evaluaciones generadas por IA.

Özer et al. (2026) proponen separar aprendizaje y evaluación en educación y, para la publicación científica, sustituir los regímenes de detección por la divulgación transparente, manteniendo la evaluación final mediante revisión humana. Las salvaguardas descritas incluyen la supervisión humana, la auditoría y trazabilidad, la responsabilidad no delegable y la divulgación del uso de IA. Su respaldo combina evaluaciones técnicas de medidas concretas con propuestas conceptuales, consensos y análisis de políticas; estos últimos no constituyen validaciones de su eficacia.

El mapa muestra una literatura muy reciente (38 de 40 fuentes cartografiadas de 2025–2026; 95,0 %), concentrada en la revisión por pares y con predominio de estudios empíricos y desarrollos y evaluaciones técnicas. La supervisión humana se menciona con frecuencia, pero su configuración pocas veces se informa de forma explícita y su evaluación en contextos editoriales reales sigue siendo limitada. Tres nodos de discusión se derivan de estos resultados. Estos resultados matizan los de las síntesis previas. Lazaros et al. (2026) sostienen, para los sistemas HITL de alto riesgo, que la automatización completa resulta

insuficiente; en la comunicación científica, sin embargo, el ~~45,0~~32,5 % de las fuentes cartografiadas evalúa tareas automatizadas sin supervisión humana durante su ejecución, de modo que la práctica evaluada va por detrás de esa recomendación. Del mismo modo, la falta de divulgación del uso de IA que Anthuvan et al. (2025) observaron en 12 de 38 estudios es coherente con la ausencia de declaraciones explícitas en muchas revistas descrita por Pérez et al. (2026).

La responsabilidad no delegable por el contenido se plantea de forma convergente en las políticas editoriales (Le Duc, 2026), en el consenso Delphi de Fettiplace et al. (2025) y en el análisis ético de Resnik y Hosseini (2025), para quienes los LLM carecen de capacidades como la consciencia, el autoconcepto, la memoria personal o la experiencia vital. La literatura de otros dominios advierte de un riesgo complementario: que la inclusión nominal de un humano en el bucle sirva sobre todo para absorber responsabilidad, el «chivo expiatorio en el bucle» (Lazaros et al., 2026); (Ranisch, 2024), y que las limitaciones de atención, competencia y capacidad de decisión reduzcan la efectividad del control humano (Constantino, 2022).

Esta literatura aborda, entre otros ámbitos, la IA médica y la administración pública; su aplicación a editores y revisores constituye una extrapolación que el corpus no ha evaluado directamente. Para que la supervisión sea significativa, se proponen autoridad para anular o interrumpir el sistema, competencia para evaluarlo y explicabilidad suficiente (Lazaros et al., 2026). En el ámbito editorial, Le Duc (2026) recomienda una competencia editorial capaz de distinguir el refinamiento lingüístico de la contribución intelectual sustantiva. Los resultados del apartado «Eje 2: sesgos cognitivos y supervisión humana» muestran que la asistencia de IA se asocia con cambios en puntuaciones y probabilidades de aceptación (Russo et al., 2025), mientras que la retroalimentación automatizada puede favorecer la actualización y la calidad de las revisiones (Thakkar et al., 2026). Sin embargo, estos resultados no permiten establecer qué condiciones de supervisión preservan un juicio humano independiente frente a recomendaciones erróneas.

Desde la teoría de sistemas sociales de Luhmann, et al. (2025) entienden la autoría científica como un acoplamiento estructural entre el sistema psíquico, la tecnología y el sistema científico, perspectiva que «desontologiza» la autoría y muestra que la relación entre individuo, tecnología y sistema social ha sido siempre dinámica.

Rogel (2025) aplica una perspectiva afín a la autoevaluación de la ciencia y sitúa la imprenta y el *preprint* entre los soportes que reconfiguraron la memoria del sistema. Watson et al. (2025) sostienen además que, si el paradigma del sistema científico prioriza la cantidad, la nueva tecnología será observada y adaptada autopoiéticamente en esa dirección, y que las soluciones centradas solo en la ética individual son insuficientes. En esta formulación, el sujeto de la adaptación autopoiética es el sistema científico, no la IA.

Esta lectura es coherente con la desalineación entre políticas editoriales y sistemas de evaluación descrita por Baciu et al. (2026) y con la advertencia de Rogel (2025) sobre las métricas calculadas en corpus contaminados, pero descansa en trabajos teóricos y en un estudio cualitativo exploratorio. La pérdida de competencias (*deskilling*) se menciona, pero no se identificaron mediciones directas de este

fenómeno en el corpus. Resnik y Hosseini (2025) advierten que confiar en la interpretación y la síntesis de la literatura realizadas por la IA, en lugar de leerla y pensarla, puede contribuir a la descualificación. Las fuentes de otros dominios también señalan el riesgo de deterioro de habilidades cuando los sistemas automatizados asumen tareas de forma sostenida (Langer et al., 2025); (Lazaros et al., 2026). Conviene distinguir la descarga cognitiva asociada a delegar una tarea de la pérdida demostrada de capacidades.

La «fricción deliberada» que se plantea aquí designa intervenciones de diseño que obliguen al revisor a registrar su juicio antes de acceder a la sugerencia de la IA. Se deriva de la recomendación de Agudo et al. (2024) de forzar un juicio explícito antes de mostrar la valoración del sistema y constituye una hipótesis de intervención para el contexto editorial cuya eficacia no ha sido validada: ninguno de los estudios de revisión por pares examinados manipuló el orden de exposición y, en el propio experimento de Agudo et al. (2024), la precisión cayó una vez recibida la recomendación errónea.

Los modelos «centauro» describen una complementariedad en la que el juicio humano y el sistema se combinan; Lazaros et al. (2026) sitúan el origen de la idea en el ajedrez. Klepikov et al. (2026) proponen, para el diagnóstico psicoeducativo, una arquitectura conceptual de «pericia experta emulada» con agentes investigador, auditor y comunicador, un agente «*escéptico*» dedicado a la falsación sistemática frente al sesgo de confirmación y registros de trazabilidad bajo supervisión *human-over-the-loop*; sus autores señalan que el diseño requiere validación a gran escala. Por su parte, Anthuvan et al. (2025) sostienen que los modelos híbridos ofrecen ventajas frente a la automatización completa, pero no aportan una estimación cuantitativa integrada de esa superioridad. La transferencia de estas arquitecturas a la revisión por pares es, por ello, una propuesta por evaluar. Los hallazgos de los tres ejes indican que la supervisión humana no puede reducirse a la presencia de una persona al final del proceso. Como propuesta, y no como resultado del corpus, la Figura 2 integra en cuatro capas las salvaguardas discutidas.

La primera sanea el manuscrito y separa las instrucciones del sistema de los datos del usuario antes de cualquier análisis, en respuesta a la «ceguera contextual» y a la inyección de texto invisible (Choi et al., 2025); (Torrielli et al., 2026). La segunda introduce fricción deliberada: el revisor registra su juicio antes de acceder a la sugerencia de la IA (Agudo et al., 2024). La tercera organiza una complementariedad de tipo «centauro» entre el investigador como ancla epistémica y una función escéptica dedicada a la falsación (Anthuvan et al., 2025); (Klepikov et al., 2026); (Lazaros et al., 2026). La cuarta documenta el proceso mediante un portafolio procesal y un registro auditable, en línea con la extensión del Manifiesto de Leiden (Thelwall, 2026).

Ninguna de estas capas se ha evaluado como conjunto y, en el único experimento disponible, registrar primero el juicio no evitó el deterioro de las decisiones ante una recomendación errónea (Agudo et al., 2024). La evidencia analizada presenta cuatro rasgos que acotan sus implicaciones. En primer lugar, muestra una marcada concentración temporal y en la revisión por pares, con numerosos estudios basados en plataformas de conferencias de aprendizaje automático como ICLR y OpenReview. En segundo lugar, predominan los estudios empíricos y los desarrollos y evaluaciones técnicas. En tercer lugar, el

examen de la conducta de editores y revisores ante la IA es todavía escaso, y no se identificaron estudios que evaluaran experimentalmente el anclaje en revisores o editores. Por último, varias de las propuestas de supervisión dependen de evidencia transferida desde dominios externos.

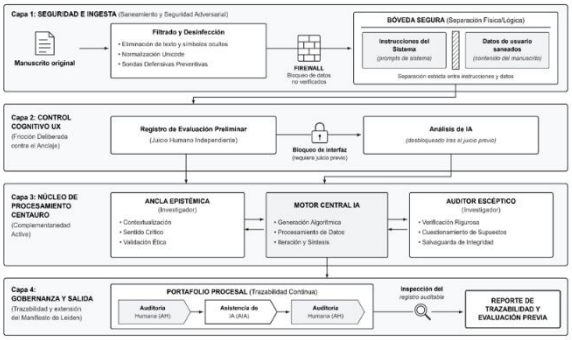


Figura 2. *Arquitectura de supervisión por capas para la evaluación de manuscritos asistida por IA (propuesta del autor).*

Fuente: Elaboración propia.

Como fortaleza, el corpus incluye análisis sobre conjuntos amplios de revisiones (Russo et al., 2025); (Thakkar et al., 2026) y experimentos con múltiples condiciones y repeticiones (Torrielli et al., 2026), aunque su alcance debe interpretarse según las tareas, los modelos y los contextos evaluados. En el subconjunto técnico y empírico directamente relacionado con la comunicación científica (n = 30), se mantuvo el predominio de la revisión por pares y la decisión editorial (24/30; 80,0 %).

Para las revistas iberoamericanas, la transferibilidad de estos hallazgos es limitada: solo tres fuentes están en español y la mayor parte de la evidencia empírica procede de conferencias de aprendizaje automático (ICLR, OpenReview). Los estudios regionales apuntan a una regulación incipiente: dos de 17 revistas colombianas mencionan la IA en sus normas (Marín y Martínez, 2025) y muchas revistas carecen de una declaración explícita sobre su uso (Pérez et al., 2026). Las salvaguardas discutidas deberían evaluarse también en revistas iberoamericanas y con manuscritos en español. Esta revisión tiene limitaciones que deben considerarse al interpretar el mapa. Primero, el cribado y la evaluación a texto completo los realizó un único revisor humano con propuestas de una herramienta de IA, y no se calculó la concordancia entre decisiones; aunque una persona ajena a la autoría resolvió las discrepancias, no hubo un segundo revisor humano independiente.

Segundo, la revisión no se registró ni dispuso de un protocolo publicado, y las cuatro reglas de agrupación se precisaron después de la selección. Los 32 registros excluidos por «ámbito no científico» y los 14 no seleccionados en las búsquedas dirigidas no se reevaluaron con la regla (b), lo que pudo infrarrepresentar la evidencia transferible. Tercero, la búsqueda se limitó a dos bases, con una cadena formulada solo en inglés pese a admitir el español, lo que pudo reducir la recuperación de literatura iberoamericana. Además, las cadenas literales de búsqueda no se conservaron, por lo que la estrategia se informa por sus bloques conceptuales. Cuarto, se excluyó la literatura gris, relevante en un campo en el que muchos resultados circulan como informes técnicos, documentos de políticas o comunicaciones no indexadas. Quinto, no se evaluó la calidad de las fuentes, por lo que el mapa describe la existencia de la evidencia y no su solidez. Por último, la fecha de corte acentúa la concentración en 2025–2026, y la rapidez del campo puede dejar el mapa desactualizado en poco tiempo.

Conclusiones

De los resultados mostrados, de su análisis y de su discusión, se obtienen las siguientes

conclusiones: La supervisión humana es la salvaguarda más invocada en la coproducción humano-IA de la comunicación científica, pero la menos demostrada: solo un tercio de las fuentes describe una configuración operativa de supervisión, mientras que las amenazas a la integridad cuentan con evidencia experimental. El «humano en el bucle» se declara, pero no se demuestra; la literatura documenta los riesgos con más rigor que la eficacia de los controles que propone para contenerlos.

La extensión del Manifiesto de Leiden propuesta recientemente, que añade el análisis costo-beneficio y el cumplimiento de los derechos de autor, ofrece una orientación normativa para la evaluación responsable cuya eficacia no ha sido validada. Para editores y comités de evaluación, se recomienda valorar la ciencia aumentada por IA considerando conjuntamente la calidad del producto, la trazabilidad del proceso y la eficacia comprobable de sus controles, de modo que la supervisión humana se acredite por la capacidad efectiva de examinar, cuestionar y corregir las decisiones asistidas por IA.

La literatura sobre la coproducción humano-IA en la comunicación científica es reciente y se centra en la revisión por pares. Los sistemas automatizados mejoran tareas específicas, pero con frecuencia operan sin supervisión humana durante la ejecución y mantienen errores, como las referencias inexistentes. La cuestión abierta es en qué fases, con qué granularidad y bajo qué condiciones debe incorporarse el juicio humano para mejorar la fiabilidad científica. Las amenazas a la integridad cuentan con evidencia experimental en escenarios controlados, especialmente la inyección de prompts y de texto invisible. También se han evaluado sondas de integridad y marcas de detección, aunque estos resultados no validan una protección integral del proceso editorial. La separación de instrucciones y datos y el saneamiento del contenido requieren evaluación específica; no se identificaron estudios sobre el efecto del procesamiento de archivos en plataformas editoriales sobre estas vulnerabilidades. La distorsión métrica es un concepto emergente sustentado en evidencia cualitativa exploratoria, sin una medición directa que permita establecer su magnitud o relación causal con el uso de IA.

Agradecimientos

Se deja constancia de un especial agradecimiento al Dr. Segundo Moisés Toapanta Toapanta, docente investigador de la Maestría en Inteligencia Artificial para la Educación de la Universidad Estatal de Milagro (UNEMI), por su escrutinio técnico, retroalimentación experta y respaldo académico durante la fase de evaluación final en la cátedra de Trabajo de Titulación II, cuyos aportes elevaron el estándar analítico de esta investigación.

Referencias Bibliográficas

Agudo, U., Liberal, K. G., Arrese, M., & Matute, H. (2024). The impact of AI errors in a human-in-the-loop process. Cognitive Research: Principles and Implications, 9, 1. https://doi.org/10.1186/s41235-023-00529-3

Anthuvan, T., Nagendra, L., Lakhani, O., Prabhuram, S., Rizvi, A. H., Maheshwari, K., & Kalra, S. (2025). Human–AI collaboration in academic writing: A narrative review and the Scholarly HI–AI Loop framework for ethical knowledge production. SSRN. https://doi.org/10.2139/ssrn.5394918

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. International Journal of Social Research Methodology, 8(1), 19–32. https://doi.org/10.1080/1364557032000119616

Baciu, C., Mohan, G., & Wilson, J. (2026). Algorithmic scholarship and academic

evaluation: Governance misalignment in the age of generative AI. Frontiers in Artificial Intelligence, 9, 1862980. https://doi.org/10.3389/frai.2026.1862980

Bagheri, F., Buscaldi, D., & Reforgiato Recupero, D. (2026). Leveraging knowledge graphs and LLMs for content-based reviewer assignment. Journal of Intelligent Information Systems, 64, 349–377. https://doi.org/10.1007/s10844-025-01004-9

Bhattacharya, S., Mazumder, A., Banerjee, A., Bandyopadhyay, C., & Nandi, S. (2025). Automated reviewer assignment process using machine learning technique. En A. Patel, N. Kesswani, M. Mishra, & P. Meher (Eds.), Advances in Machine Learning and Big Data Analytics I (ICMLBDA 2023) (pp. 87–99). Springer. https://doi.org/10.1007/978-3-031-51338-1_7

Carpenter, J. M., Corvillón, A., & Shah, N. B. (2025). Enhancing peer review in astronomy: A machine learning and optimization approach to reviewer assignments for ALMA. Publications of the Astronomical Society of the Pacific, 137, 34501. https://doi.org/10.1088/1538-3873/adb5c1

Choi, B., Jun, T. J., Sung, J. W., Park, I., Lee, J. M., Cho, S. I., Park, H. J., Lee, R. W., & Suh, J. (2025). Invisible text injection: The trojan horse of AI-assisted medical peer review. medRxiv. https://doi.org/10.1101/2025.07.24.25332148

Constantino, J. E. (2022). Exploring Article 14 of the EU AI proposal: Accountability challenges of the human in the loop when supervising high-risk AI systems in public administration. https://doi.org/10.2139/ssrn.4254940

Committee on Publication Ethics. (2023). Authorship and AI tools: COPE position statement. https://doi.org/10.24318/cCVRZBms

Daoudi, M. (2025). Ethical limits and suggestions for improving the use of AI in scientific research, academic publishing, and the peer review process, based on deontological and consequentialist viewpoints. Discover Education, 4, 241. https://doi.org/10.1007/s44217-025-00696-z

Fettiplace, M. R., Bhatia, A., Chen, Y., Orebaugh, S. L., Gofeld, M., Gabriel, R. A., Sessler, D. I., Lonsdale, H., Bungart, B., Cheng, C. P., Burnett, G. W., Wiles, M., Coppens, S., Schreiber, K. L., Volk, T., Wu, C. L., Mariano, E. R., & Ip, V. H. Y. (2025). Recommendations for disclosure of artificial intelligence in scientific writing and publishing: A regional anesthesia and pain medicine modified Delphi study. Regional Anesthesia & Pain Medicine, 51(6), 603–610. https://doi.org/10.1136/rapm-2025-106852

Guerard, G., Djebali, S., Hanus, M., & Zinenberg, M. K. (2026). Research LLM: Using large language models as a research companion. Tehnički Glasnik, 20(1), 118–125. https://doi.org/10.31803/tg-20250827131608

Idahl, M., & Ahmadi, Z. (2025). OpenReviewer: A specialized large language model for generating critical scientific paper reviews. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), 550–562. https://doi.org/10.18653/v1/2025.naacl-demo.44

Isik, M., & Guleryuzlu, S. (2026). A large language model pipeline for automated citation quality scoring across engineering journal quartiles with expert validation. Scientific Reports, 16, 23628. https://doi.org/10.1038/s41598-026-60947-3

Klepikov, O. E., Pozdnyakova, E. M., & Sizov, Y. O. (2026). From measurement to understanding: An architecture of emulated expertise for LLM-diagnostics in education. Psychology. Journal of the Higher School of Economics, 23(2), 263–274. https://doi.org/10.17323/1813-8918-2026-2-263-274

Kruzenshtern, A., Dodonov, V., & Chechurin, L. (2026). Domain agnostic features for robust novelty assessment of scientific

publication. Scientometrics, 131, 4683–4709. https://doi.org/10.1007/s11192-026-05687-x

Langer, M., Baum, K., & Schlicker, N. (2025). Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. Minds and Machines, 35, 1. https://doi.org/10.1007/s11023-024-09701-0

Lazaros, K., Vrahatis, A. G., & Kotsiantis, S. (2026). Human-in-the-loop artificial intelligence: A systematic review of concepts, methods, and applications. Entropy, 28(4), 377. https://doi.org/10.3390/e28040377

Le Duc, A. (2026). Codification of scholarly values in academic publishing: AI governance and institutional tensions. Prajñā Vihāra, 27(2), 5–34. https://doi.org/10.59865/prajn.2026.8

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI, 1(8). https://doi.org/10.1056/AIoa2400196

Majadly, M., & Last, M. (2025). Leveraging peer-review aspects for extractive and abstractive summarization of scientific articles. International Journal of Data Science and Analytics, 19, 537–550. https://doi.org/10.1007/s41060-024-00665-z

Marín, F., & Martínez, P. (2025). Desafíos y estrategias en los procesos de normalización de revistas científicas en el área educativa. Revista Científica General José María Córdova, 23(49), 139–166. https://doi.org/10.21830/19006586.1406

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when deciding between a systematic or scoping review approach. BMC Medical Research Methodology, 18, 143. https://doi.org/10.1186/s12874-018-0611-x

Nathani, K. R., Nathani, A. M., Delawan, M., Safdar, A., & Bydon, M. (2025). Can artificial intelligence write science? A comparative analysis of human-written and artificial intelligence-generated scientific writings. Journal of Neurosurgery: Spine, 43(6), 767–772. https://doi.org/10.3171/2025.4.SPINE25519

Özer, M., Perc, M., & Özçelik, H. T. (2026). Human agency and epistemic authority under generative artificial intelligence. Open Praxis, 18(3), 495–505. https://doi.org/10.55982/openpraxis.18.3.1149

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ, 372, n160. https://doi.org/10.1136/bmj.n160

Pandey, C. S., Pandey, S., Pandey, T., Pandey, S., Pandey, H., & Mishra, P. (2025). A comparative study of machine learning, natural language processing and hybrid models for academic paper acceptance prediction: From reviews to decisions. DESIDOC Journal of Library & Information Technology, 45(5), 439–447. https://doi.org/10.14429/djlit.21138

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Patel, R., Shultz, C., Ollivier, M., & Wascher, D. C. (2026). Artificial intelligence cannot replace peer reviewers but may help editors triage: A comparative analysis of a large language model and human reviewer recommendations at the American Journal of Sports Medicine. The American Journal of Sports Medicine, 54(11), 2800–2807. https://doi.org/10.1177/03635465261463006

Pendyala, V. S., Kamdar, K., & Mulchandani, K. (2025). Automated research review support using machine learning, large language models, and natural language

processing. Electronics, 14(2), 256. https://doi.org/10.3390/electronics14020256

Pérez, A., Rodríguez-, C., Guillén, M., & Pita Mera, M. V. (2026). Transformación de procesos y políticas editoriales de las revistas científicas frente al uso de inteligencia artificial. Revista de Ciencias Sociales (Ve), 32(1), 278–305. https://www.redalyc.org/articulo.oa?id=28085564020

Peters, M, Godfrey, C., McInerney, P., Khalil, H., Larsen, P., Marnie, C., Pollock, D., Tricco, A. C., & Munn, Z. (2022). Best practice guidance and reporting items for the development of scoping review protocols. JBI Evidence Synthesis, 20(4), 953–968. https://doi.org/10.11124/JBIES-21-00242

Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. JBI Evidence Synthesis, 18(10), 2119–2126. https://doi.org/10.11124/JBIES-20-00167

Pollock, D., Peters, M. D. J., Khalil, H., McInerney, P., Alexander, L., Tricco, A. C., Evans, C., de Moraes, E. B., Godfrey, C. M., Pieper, D., Saran, A., Stern, C., & Munn, Z. (2023). Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evidence Synthesis, 21(3), 520–532. https://doi.org/10.11124/JBIES-22-00123

Ranisch, R. (2024). Scapegoat-in-the-loop? Human control over medical AI and the (mis)attribution of responsibility. The American Journal of Bioethics, 24(9), 116–117. https://doi.org/10.1080/15265161.2024.2377137

Rao, VKumar, A., Lakkaraju, H., & Shah, N. B. (2025). Detecting LLM-generated peer reviews. PLOS ONE, 20(9), e0331871. https://doi.org/10.1371/journal.pone.0331871

Resnik, D. B., & Hosseini, M. (2025). The ethics of using artificial intelligence in scientific research: New guidance needed for a new tool. AI and Ethics, 5, 1499–1521. https://doi.org/10.1007/s43681-024-00493-8

Rogel, R. (2025). De la imprenta al algoritmo: Reconfiguración sistémica de la autoevaluación científica. MAD. Revista del Magíster en Análisis Sistémico Aplicado a la Sociedad, (53), 80–107. https://doi.org/10.5354/0719-0527.2025.82301

Russo, G., Horta, M., Davidson, T. R., Veselovsky, V., & West, R. (2025). The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. Proceedings of the ACM on Human-Computer Interaction, 9(7), 1–28. https://doi.org/10.1145/3757667

Sharma, N. K., & Sarode, S. C. (2026). Human-assisted traditional peer review intelligence vs artificial peer review intelligence: Perspectives on future disruptive sciences. Discover Artificial Intelligence, 6, 246. https://doi.org/10.1007/s44163-026-01013-9

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Sun, Q., Liu, D., Dong, H., & Li, S. (2026). Automatically investigating scientific discussions in peer review reports based on conformity score metrics. Scientometrics, 131, 59–92. https://doi.org/10.1007/s11192-026-05539-8

Thakkar, N., Yuksekgonul, M., Silberg, J., Garg, A., Peng, N., Sha, F., Yu, R., Vondrick, C., & Zou, J. (2026). A large-scale randomized study of large language model feedback in peer review. Nature Machine Intelligence, 8, 326–336. https://doi.org/10.1038/s42256-026-01188-x

Thelwall, M. (2025). Can smaller large language models evaluate research quality? Malaysian Journal of Library & Information Science, 30(2), 66–81. https://doi.org/10.22452/mjlis.vol30no2.4

Thelwall, M. (2026). Large language models and responsible research evaluation: An extension of the Leiden Manifesto.

Scientometrics. https://doi.org/10.1007/s11192-026-05552-x

Thelwall, M., & Mohammadi, E. (2026). Can small and reasoning large language models score journal articles for research quality and do averaging and few-shot help? Scientometrics, 131(5), 3525–3546. https://doi.org/10.1007/s11192-026-05585-2

Thelwall, M., & Yaghi, A. (2025). Evaluating the predictive capacity of ChatGPT for academic peer review outcomes across multiple platforms. Scientometrics, 130, 5285–5307. https://doi.org/10.1007/s11192-025-05287-1

Torrielli, F., Locci, S., Rapp, A., & Di Caro, L. (2026). Exploiting large language models in peer review: Indirect prompt injection attacks and integrity probes. Scientometrics, 131, 4889–4951. https://doi.org/10.1007/s11192-026-05695-x

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., & Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Annals of Internal Medicine, 169(7), 467–473. https://doi.org/10.7326/M18-0850

Watson, S., Brezovec, E., & Romic, J. (2025). The role of generative AI in academic and scientific authorship: An autopoietic perspective. AI & Society, 40, 3225–3235. https://doi.org/10.1007/s00146-024-02174-w

Xu, X., Yan, J., Nwachukwu, G., Shan, H., Kruger, U., & Wang, G. (2026). Artificial intelligence-aided assignment of journal submissions to associate editors—A feasibility study on IEEE Transactions on Medical Imaging. Visual Computing for Industry, Biomedicine, and Art, 9, 1. https://doi.org/10.1186/s42492-025-00212-y

Yuan, C., Wei, T., Li, C., Yi, X., Liu, S., Zhang, Z., Cai, Y., & Du, X. (2026). PaperOrchestrator: An LLM-orchestrated multi-agent pipeline for automated end-to-end scientific paper writing. Journal of King Saud University – Computer and Information Sciences, 38(5), 296. https://doi.org/10.1007/s44443-026-00708-4

Esta obra está bajo una licencia de Creative Commons Reconocimiento-No Comercial 4.0 Internacional. Copyright © Andrés Guillermo Chérrez Ramírez.

Declaraciones éticas y editoriales del artículo
Contribución de los autores (Taxonomía CRediT) Andrés Guillermo Chérrez Ramírez.: conceptualización de la investigación, diseño metodológico, desarrollo del proceso investigativo, análisis formal de los datos, redacción del borrador original del manuscrito, revisión crítica del contenido científico y supervisión general del estudio.
Declaración de conflicto de intereses Los autores declaran que no existe conflicto de intereses en relación con la investigación presentada, la autoría del manuscrito ni la publicación del presente artículo.
Declaración de financiamiento La presente investigación no recibió financiamiento específico de agencias públicas, comerciales o de organizaciones sin fines de lucro. En caso de existir financiamiento institucional o externo, este deberá ser declarado explícitamente por los autores en esta sección.
Declaración del editor El editor responsable certifica que el proceso editorial del presente artículo se desarrolló conforme a los principios de integridad científica, transparencia y buenas prácticas editoriales. El manuscrito fue sometido a un proceso de evaluación mediante revisión por pares doble ciego, garantizando la confidencialidad de la identidad de los autores y revisores durante todo el proceso de dictamen académico. Asimismo, el editor declara que el artículo cumple con los criterios científicos, metodológicos y éticos establecidos por la revista.
Declaración de los revisores Los revisores externos que participaron en la evaluación del presente manuscrito declaran haber realizado el proceso de revisión de manera objetiva, independiente y confidencial. Asimismo, manifiestan que no mantienen conflictos de interés con los autores ni con la investigación evaluada, y que sus observaciones y recomendaciones se fundamentan exclusivamente en criterios científicos, metodológicos y académicos.
Declaración ética de la investigación Los autores declaran que la investigación se desarrolló respetando los principios éticos de la investigación científica, garantizando la confidencialidad de los datos y el respeto a los participantes del estudio. En los casos en que la investigación involucre seres humanos, los procedimientos deben ajustarse a los principios éticos establecidos en la Declaración de Helsinki y a las normativas institucionales correspondientes.
Declaración sobre el uso de inteligencia artificial Los autores declaran que el uso de herramientas de inteligencia artificial, en caso de haberse utilizado durante el proceso de investigación o redacción del manuscrito, se realizó únicamente como apoyo técnico para mejorar la claridad del lenguaje o el análisis de información, manteniendo siempre la responsabilidad intelectual sobre el contenido del artículo. Las herramientas de inteligencia artificial no fueron utilizadas como autoras del manuscrito ni sustituyen la responsabilidad académica de los investigadores.
Disponibilidad de datos Los datos que respaldan los resultados de esta investigación estarán disponibles previa solicitud razonable al autor de correspondencia, respetando las normas éticas y de confidencialidad establecidas por la investigación.