Riesgos del uso de Inteligencia Artificial Generativa en comunicación e investigación científica

William Mauricio Rojas Contreras
Olga Liliana Rojas Contreras
Eliana Caterine Mojica Acevedo

UNIVERSIDAD DE PAMPLONA

Riesgos del uso de Inteligencia Artificial Generativa en comunicación e investigación científica

Riesgos del uso de Inteligencia Artificial Generativa en comunicación e investigación científica / William Mauricio Rojas Contreras, Olga Liliana Rojas Contreras, Eliana Caterine Mojica Acevedo -- Pamplona: Universidad de Pamplona. 2025.

122 p. 17 cm x 24 cm

ISBN (Digital): 978-628-7656-87-1

Sede Principal Pamplona, Km 1 Vía Bucaramanga-
Ciudad Universitaria. Norte de Santander, Colombia.
www.unipamplona.edu.co
Teléfono: 6075685303

Riesgos del uso de Inteligencia Artificial Generativa en comunicación e investigación científica

William Mauricio Rojas Contreras
Olga Liliana Rojas Contreras
Eliana Caterine Mojica Acevedo

ISBN (Digital): 978-628-7656-87-1
Primera edición diciembre de 2025
Colección Ciencia e Innovación

Rector: Ivaldo Torres Chávez Ph.D
Vicerrector de Investigaciones: Aldo Pardo García Ph.D

Jefe Sello Editorial Unipamplona: Caterine Mojica Acevedo
Diseño y diagramación: Laura Angelica Buitrago Quintero

Cubierta ilustrada con apoyo de Inteligencia Artificial

Riesgos del uso de Inteligencia Artificial Generativa en comunicación e investigación científica

William Mauricio Rojas Contreras
Olga Liliana Rojas Contreras
Eliana Caterine Mojica Acevedo

Formando nuevas generaciones con sello de excelencia comprometidos con la transformación social de las regiones y un país en paz

SELLO EDITORIAL
UNIVERSIDAD DE PAMPLONA

Vicerrectoría de INVESTIGACIONES UNIPAMPLONA

PRÓLOGO

La Inteligencia Artificial Generativa (IAG) ha transformado radicalmente la manera en que se crea, difunde y consume información en la comunicación e investigación científica. Sin embargo, esta revolución tecnológica ha traído consigo desafíos significativos, particularmente en lo que respecta a la proliferación de desinformación. Los modelos de IA generativa, al producir contenido sin verificación humana, pueden introducir errores, sesgos y alucinaciones que comprometen la integridad del conocimiento científico. Ante este panorama, se vuelve imperativo analizar los riesgos que emergen del uso de la IA en los procesos de comunicación, investigación y explorar estrategias para su mitigación.

Este libro, Riesgos en el uso de la IAG en la comunicación e investigación científica, tiene como propósito examinar de manera rigurosa los diversos riesgos asociados al uso de la inteligencia artificial en la generación y difusión de información. Se inicia con una revisión de la literatura sobre los dominios de riesgos de la IA en general, destacando su impacto en la comunicación e investigación científica. Posteriormente, se centra en el dominio específico de los riesgos de desinformación, identificando sus principales causas, manifestaciones y consecuencias en la comunidad académica y en la sociedad en general. Finalmente, se presentan alternativas para mitigar estos riesgos, con un énfasis particular en el enfoque Human-In-The-Loop (HITL), el cual propone la intervención humana como un elemento fundamental para garantizar la calidad y veracidad del conocimiento generado con IA.

El presente trabajo no solo busca documentar el estado actual del problema, sino también ofrecer un marco analítico que contribuya al desarrollo de estrategias efectivas para enfrentar la desinformación en la comunicación e investigación científica. Se espera que este libro sea de utilidad para investigadores,

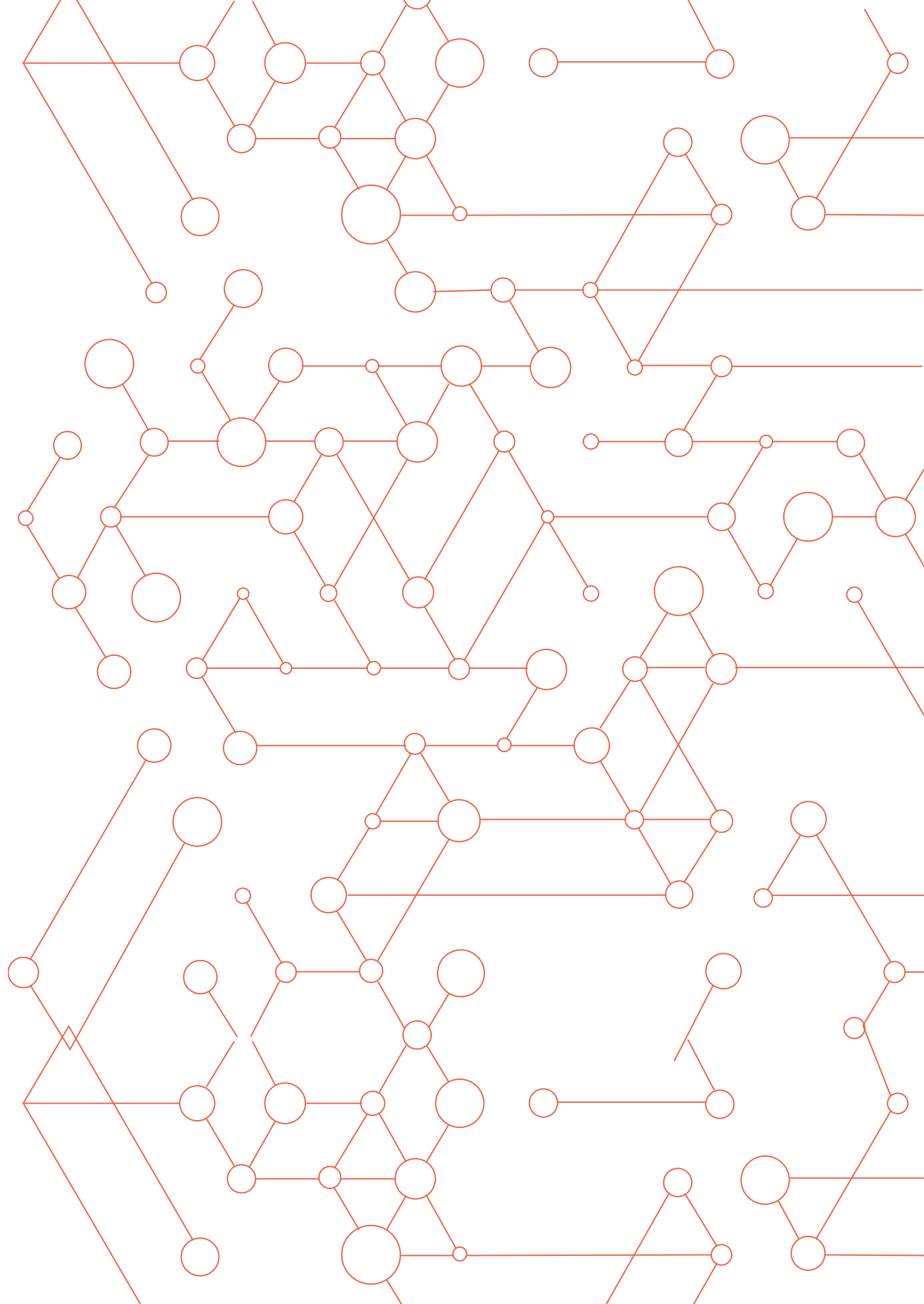
educadores, comunicadores y responsables de políticas que buscan comprender los riesgos inherentes al uso de la IA y promover un uso ético y responsable de estas tecnologías.

En un contexto donde la inteligencia artificial continuará evolucionando y desempeñando un papel central en la producción de conocimiento, la adopción de enfoques críticos y metodologías de mitigación será esencial para preservar la integridad del discurso científico. Esta obra invita al lector a reflexionar sobre los desafíos actuales y futuros, fomentando una discusión informada que contribuya a un ecosistema de comunicación más transparente, confiable y basado en principios éticos sólidos.

RESUMEN

El desarrollo y la integración de la inteligencia artificial (IA) han planteado importantes desafíos éticos, especialmente en el dominio de la desinformación. Este trabajo de investigación aborda los riesgos específicos asociados a la IA generativa en los procesos de comunicación científica, respondiendo a la pregunta: ¿Cuáles son los subdominios del dominio de desinformación que se generan en la comunicación e investigación científica por el uso de la inteligencia artificial? A través de una revisión sistemática de literatura, se identificaron los principales dominios de riesgos de la IA, tales como discriminación y toxicidad, privacidad y seguridad, desinformación, actores maliciosos, interacción humano-computadora, socioeconómicos y ambientales, y fallos de los sistemas de IA. El dominio de desinformación se desglosa en los subdominios de Información falsa o engañosa y Contaminación del ecosistema informativo y pérdida de la realidad consensuada. Los riesgos más frecuentes identificados en este dominio incluyen el sesgo de datos y la desinformación intencional. La investigación contribuye al campo de la ética de la IA, proponiendo recomendaciones para un uso ético y responsable de la IA en la ciencia. Futuros estudios deberán centrarse en el desarrollo de estrategias de mitigación de riesgos y en abordar las brechas identificadas en el dominio de la desinformación.

Palabras Claves: Alucinación, desinformación, ética de la IA, inteligencia artificial, sesgo de datos.



ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La aparición reciente de sistemas de inteligencia artificial ha impulsado la innovación, optimización y transformación de diversos campos disciplinarios, incluyendo la salud (Rogers et al., 2021; Dhawan & Kumar, 2024; Williamson & Prybutok, 2024), la educación (Niveditha et al., 2024; Abimbola et al., 2024) y la comunicación (Al-Rashed, 2024; Ghani et al., 2023; Samuel-Okon, 2024). En particular, las inteligencias artificiales generativas y las herramientas impulsadas por IA han permitido agilizar la generación de conocimiento en comparación con los métodos tradicionales, mejorando los tiempos de respuesta. No obstante, este enfoque ha tendido a centrarse en los aspectos instrumentales, sin prestar la debida atención a la verificación y validación del conocimiento generado.

Si bien la integración de herramientas de IA ha añadido valor a múltiples disciplinas, también ha planteado preocupaciones éticas importantes. El uso de la inteligencia artificial conlleva riesgos asociados, tanto en la forma en que se produce y difunde el conocimiento, como en la interpretación de este por parte de los usuarios. Estudios recientes (Manoj & Sahil, 2025) han destacado varios desafíos éticos, incluyendo la privacidad, la seguridad, el sesgo, la equidad, la transparencia y la fiabilidad de los sistemas de IA. Otros trabajos, como el de Jiao et al. (2024), se han enfocado en los retos éticos asociados a los grandes modelos de lenguaje (LLM), como la alucinación, la responsabilidad verificable y las dificultades inherentes a la censura de decodificación. Slattery et al. (2024) aportan una clasificación más exhaustiva de los riesgos, agrupándolos en siete dominios clave: discriminación y toxicidad, privacidad y seguridad, desinformación, actores maliciosos, interacción humano-computadora, impactos socioeconómicos y ambientales, y seguridad de los sistemas de IA.

Este libro se centra en uno de los dominios menos explorados: la desinformación. A pesar de la creciente preocupación por la difusión de información errónea en entornos digitales, el análisis de los riesgos asociados a la desinformación generada por IA ha sido limitado. Slattery et al. (2024) identifican dos subdominios dentro de este riesgo: (1) la información falsa o engañosa, que incluye el fenómeno de la alucinación en los LLM, y (2) la contaminación del ecosistema informativo, que aborda la pérdida de una realidad consensuada en la sociedad.

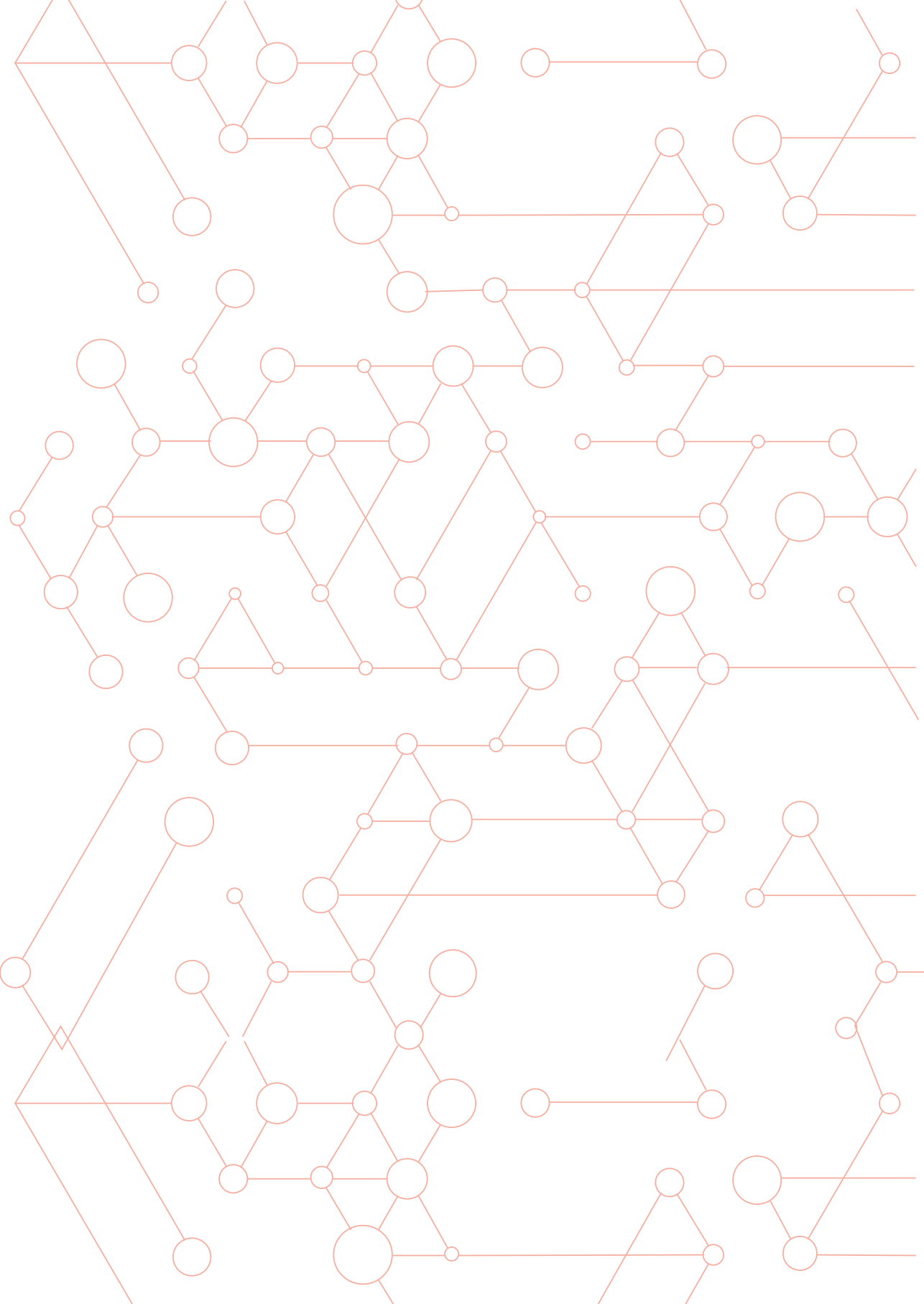
El riesgo de alucinación se manifiesta cuando los sistemas de IA producen contenido que parece plausible pero que no se basa en hechos reales. Este tipo de error puede erosionar la confianza en el conocimiento generado por IA y tener efectos graves en la comunicación científica. Del mismo modo, la falta de contexto en la información generada puede conducir a malentendidos, impactando la credibilidad de las fuentes y dificultando la toma de decisiones informadas.

La presente investigación tiene como objetivo abordar las brechas identificadas en el análisis del dominio de desinformación, con el fin de proporcionar una mejor comprensión de los riesgos asociados y proponer estrategias de mitigación específicas para su aplicación en la comunicación científica. En particular, se plantea la siguiente pregunta de investigación:

P1: ¿Cuáles son los subdominios del dominio de desinformación que se generan en la comunicación e investigación científica por el uso de la inteligencia artificial?

CAPÍTULO I

EL PROBLEMA DE LA DESINFORMACIÓN EN LA COMUNICACIÓN E INVESTIGACIÓN CIENTÍFICA



CAPÍTULO I

EL PROBLEMA DE LA DESINFORMACIÓN EN LA COMUNICACIÓN E INVESTIGACIÓN CIENTÍFICA

1.1 Introducción

Las inteligencias artificiales generativas y las herramientas impulsadas por inteligencia artificial han mejorado la eficiencia en los procesos de comunicación académica, facilitando tareas como la generación de textos e imágenes, la corrección de redacción, gramática y estilo, la revisión sistemática de literatura, la estructuración disciplinada de artículos de investigación y la difusión del conocimiento. Sin embargo, el uso de estas tecnologías en la comunicación académica también exige la planificación de estrategias de mitigación para diversos riesgos éticos, entre ellos, el riesgo de desinformación. La generación de contenido académico y de investigación mediante IA puede introducir alucinaciones, información inexacta o no respaldada científicamente, así como amplificar sesgos existentes, lo que compromete la integridad, confiabilidad, veracidad y uso ético del conocimiento académico.

Esta sección trata el problema de la desinformación como uno de los desafíos centrales en el uso de la inteligencia artificial generativa y herramientas impulsadas por inteligencia artificial en los procesos de comunicación académica. Se analiza como la Inteligencia Artificial Generativa - IAG, puede generar conocimiento, amplificar y reutilizar desinformación en contextos de comunicación académica, lo cual impacta en la integridad, veracidad y confianza en la academia y la investigación.

1.2 Contextualización del Problema

El uso de inteligencias artificiales generativas y herramientas basadas en inteligencia artificial en los procesos de comunicación académica ha aumentado significativamente con la aparición de los grandes modelos de lenguaje (LLM, por sus siglas en inglés). Estas tecnologías permiten a académicos y estudiantes generar textos, imágenes y videos de manera más eficiente a partir de indicaciones o prompts. Sin embargo, la generación de conocimiento mediante inteligencia artificial generativa también plantea riesgos, particularmente en relación con la desinformación, la precisión y la veracidad del contenido producido.

El riesgo de desinformación en la comunicación académica puede manifestarse de diferentes maneras, iniciando desde la generación de contenido sin fundamentación científica, pasando por la generación de citas y referencias no existentes, hasta la generación de alucinaciones. La capacidad de la IAG para generar contenidos fiables dificulta la identificación de alucinaciones y aumenta el riesgo de que la información generada se reutilice en artefactos académicos sin verificación y validación de este tipo de conocimiento.

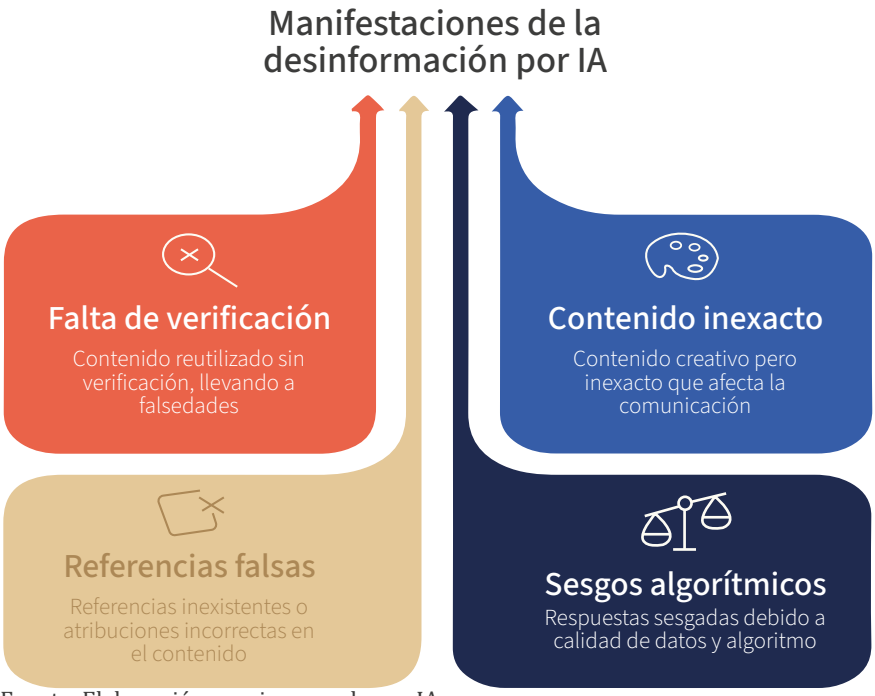
En forma complementaria, los sesgos algorítmicos pueden impactar la generación de conocimiento, pueden generar conocimiento erróneo derivado del sesgo que se incorporan en el diseño de los algoritmos que soportan la implementación de los sistemas de IA. También, este tipo de sesgo puede incorporar escenarios en los cuales se omiten orientaciones relevantes en la generación de conocimiento. Desde este punto de vista, la desinformación no solo tiene implicaciones de calidad en la comunicación académica sino también tiene consideraciones de orden ético.

1.3 Manifestaciones de la Desinformación en la Comunicación Académica

La desinformación generada por la IA se puede manifestar en la comunicación académica de diferentes maneras, entre ellas las enunciadas en la Figura 1. En forma específica, en la Figura 1 se visualizan las manifestaciones de la desinformación en la

comunicación por impacto de la IA como son la generación de contenido inexacto, manipulación de datos y sesgos algorítmicos, fabricación de referencias, atribución errónea, ausencia de verificación y validación.

Figura 1
Manifestaciones de la desinformación en la comunicación académica



Fuente: Elaboración propia apoyada con IA.

1.3.1 Generación de contenidos con falta de precisión y alta creatividad

Las inteligencias artificiales generativas pueden producir conocimiento con altos niveles de creatividad, lo cual puede llevar a contenidos con falta de precisión, impactando los procesos de comunicación académica si estos contenidos se reutilizan y difunden sin verificación y validación del conocimiento.

1.3.2 Manipulación de Datos y Sesgos Algorítmicos

Las inteligencias artificiales generativas, en sus etapas de diseño, implementación y entrenamiento, pueden presentar sesgos derivados de la calidad, completitud e integridad de los datos utilizados para su entrenamiento. Además, el diseño de los algoritmos puede incorporar sesgos influenciados por la filosofía de implementación de sus desarrolladores. Estos factores pueden conducir a la generación de respuestas incorrectas o parciales por parte de las inteligencias artificiales generativas.

1.3.4 Fabricación de Referencias, Atribución Errónea

Las inteligencias artificiales generativas pueden producir contenido con referencias que, al ser verificadas, resultan inexistentes. Asimismo, pueden atribuir referencias a segmentos de texto generado que no corresponden con los autores citados.

1.3.5 Ausencia de Verificación y Validación del Conocimiento

La falta de aplicación de estrategias de verificación y validación del conocimiento generado por las inteligencias artificiales generativas puede llevar a la reutilización de su contenido en artículos de comunicación que carecen de precisión, presentan un alto grado de creatividad y pueden resultar potencialmente falsos.

1.4 Impacto Ético de la Desinformación en los Procesos de Comunicación Académica

La desinformación en los procesos de comunicación académica puede traer consecuencias críticas de orden ético, como la pérdida de credibilidad y confianza en las instituciones educativas. En forma complementaria, la desinformación puede impactar la calidad de los resultados de investigación y potencialmente aumentar la responsabilidad de los autores que utilizan en sus artículos contenido generado con inteligencia artificial generativa.

1.5 Justificación de la Investigación

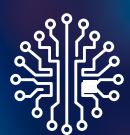
Debido a la creciente adopción de la inteligencia artificial generativa y de herramientas basadas en IA en los procesos de comunicación académica, resulta fundamental abordar el problema de la desinformación desde una perspectiva ética. La falta de políticas, normativas y mecanismos claros para la verificación y validación del conocimiento generado por estas tecnologías resalta la necesidad de investigar la desinformación y desarrollar métodos para mitigar el impacto del uso indebido del conocimiento producido por las inteligencias artificiales generativas.

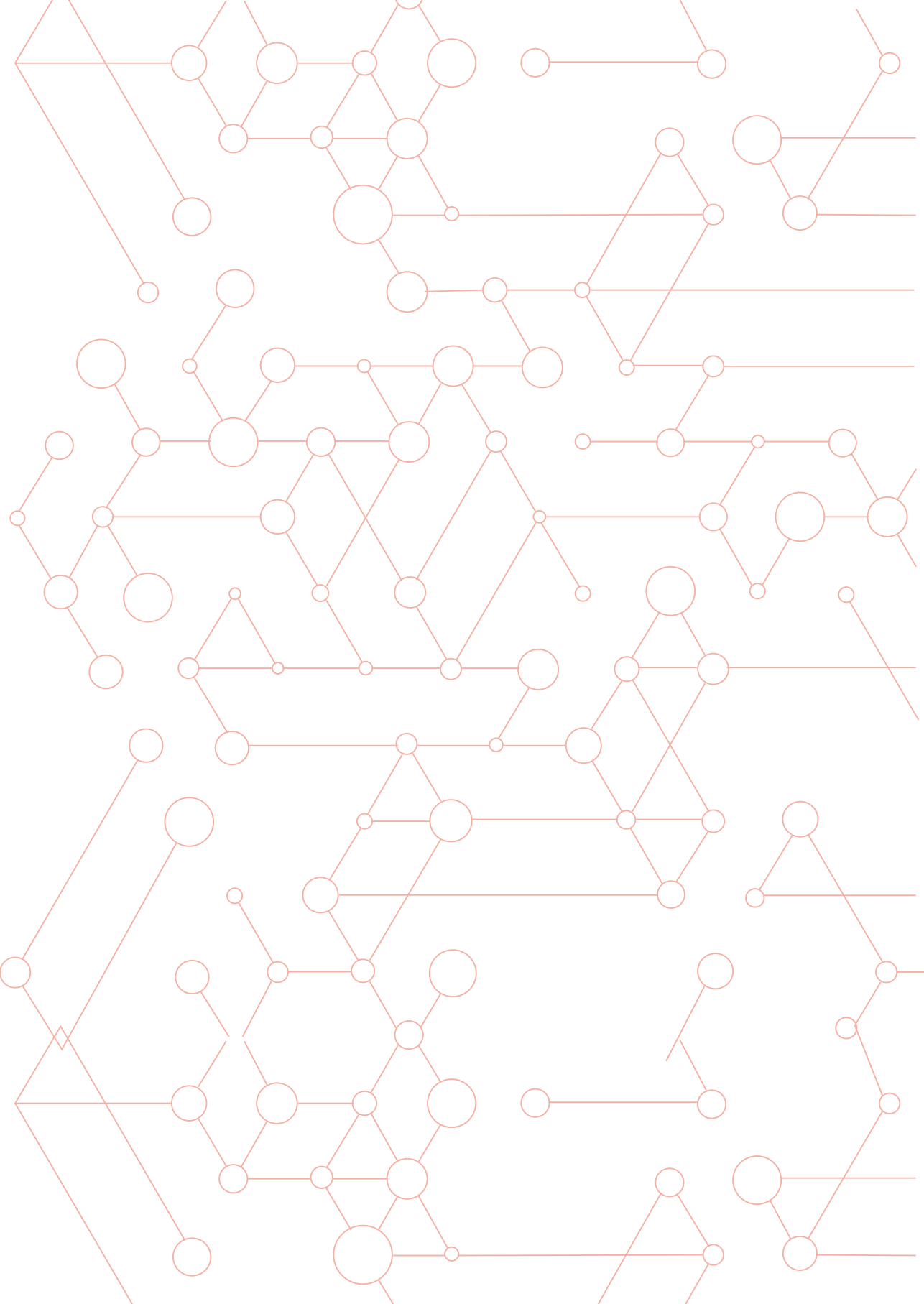
Esta investigación busca contribuir al debate sobre los aspectos éticos del uso de la inteligencia artificial en la comunicación académica, con un enfoque en el riesgo de desinformación. En particular, analiza las instancias de riesgo dentro de este dominio y propone estrategias y recomendaciones para preservar la integridad del conocimiento científico.

La desinformación producida por la inteligencia artificial generativa representa un gran reto para la comunicación académica. Este trabajo plantea la necesidad de un análisis ético profundo y de estrategias concretas para mitigar sus riesgos, con el fin de proteger la integridad del conocimiento.

CAPÍTULO II

ELEMENTOS TEÓRICOS RELACIONADOS CON LA ÉTICA DE LA INTELIGENCIA ARTIFICIAL EN LOS PROCESOS DE COMUNICACIÓN E INVESTIGACIÓN CIENTÍFICA





CAPÍTULO II

ELEMENTOS TEÓRICOS RELACIONADOS CON LA ÉTICA DE LA INTELIGENCIA ARTIFICIAL EN LOS PROCESOS DE COMUNICACIÓN E INVESTIGACIÓN CIENTÍFICA

2.1 Fundamentos de la Ética en la Inteligencia Artificial

La implementación de sistemas de IA y el uso de herramientas impulsadas por esta tecnología en los procesos de comunicación académica han propiciado avances significativos. No obstante, también han generado una serie de desafíos éticos que requieren un análisis riguroso y la formulación de planes de mitigación. En este capítulo se examinan los fundamentos éticos del uso de la IA en la comunicación científica, los riesgos asociados y el impacto social derivado de su aplicación, así como las estrategias para mitigar dichos riesgos.

2.1.1 Principios Éticos Fundamentales en IA

La inteligencia artificial plantea una serie de desafíos éticos que requieren la definición de principios fundamentales para su desarrollo, aplicación y uso del conocimiento generado. Entre estos principios se encuentran la beneficencia, la no maleficencia, la autonomía, la transparencia, la justicia, la equidad, la explicabilidad, la seguridad, la responsabilidad y la privacidad. Su propósito es garantizar que los sistemas de IA y el conocimiento derivado de ellos se utilicen de manera justa y respetuosa con los derechos humanos.

2.1.2 Marco Normativo y Regulatorio

La rápida evolución en el diseño, desarrollo, implementación y uso del conocimiento generado por los sistemas de IA, han llevado a diversos organismos internacionales y gobiernos nacionales a establecer marcos regulatorios específicos. En el ámbito internacional, destacan las iniciativas de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), la Organización para la Cooperación y el Desarrollo Económico (OCDE) y la Unión Europea. En Colombia, el gobierno ha puesto recientemente a disposición de la comunidad la Política Nacional de Inteligencia Artificial.

2.1.3 Responsabilidad, Integridad de la Información, Explicabilidad y Transparencia Algorítmica

La responsabilidad, la explicabilidad y la transparencia en el diseño de los algoritmos que sustentan los sistemas de IA, así como la garantía de la integridad de la información y del conocimiento generado por las inteligencias artificiales generativas, son aspectos clave para minimizar los riesgos éticos. Estos elementos contribuyen al control de los sistemas de IA y al uso responsable del conocimiento que producen.

2.2 Riesgos Éticos en IA

El uso de la IA en los procesos de comunicación científica implica una serie de riesgos, los cuales pueden afectar tanto a individuos como a la sociedad en general. A continuación, se analizan los principales riesgos, su valoración y el impacto social generado.

2.2.1 Clasificación de Riesgos Éticos en IA

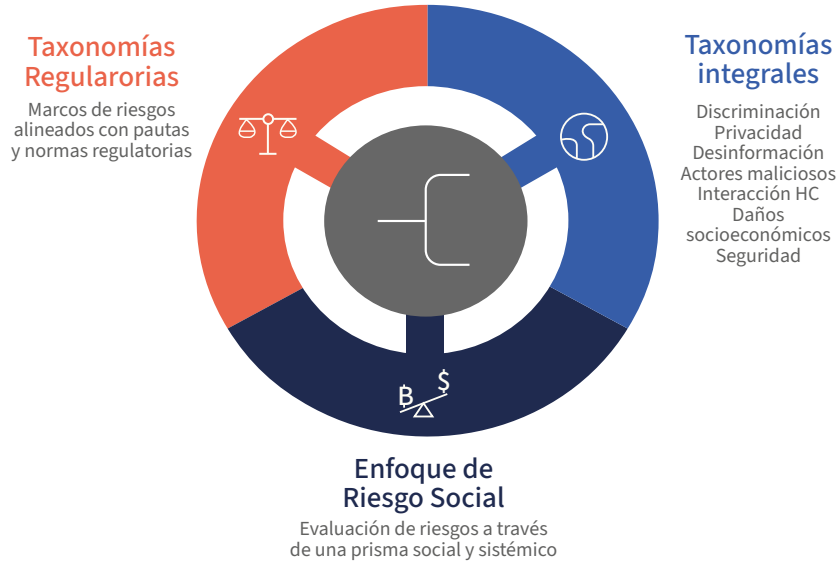
Los riesgos éticos de la IA inicialmente se pueden clasificar en riesgos técnicos, riesgos sociales y riesgos legales. Los riesgos técnicos, son potencialmente manifestados por errores en los algoritmos, sesgos en los datos y vulnerabilidades de seguridad; los riesgos sociales, tienen por alcance temas relacionados con la desinformación, la manipulación de la opinión pública y la

transgresión de la privacidad de los usuarios de la IA; los riesgos legales, están asociados con el incumplimiento de normas y la falta de gestión en la asignación de responsabilidades.

Por otro lado, la revisión de la literatura identifica enfoques modernos para la clasificación de los riesgos. Entre ellos se destacan las taxonomías integrales (Slattery et al., 2024; Zeng et al., 2024), el enfoque de riesgo social y sistémico (Critch & Russell, 2023) y las taxonomías asociadas a la regulación (Golpayegani et al., 2022; Novelli et al., 2024).

En el contexto de esta investigación, se focaliza el estudio en las taxonomías integrales, las cuales identifican los siguientes dominios de riesgo: discriminación y toxicidad, privacidad y seguridad, desinformación, actores maliciosos y uso indebido, interacción humano-computadora, daños socioeconómicos y ambientales, y seguridad de los sistemas de IA, como se ilustra en la figura 2. En forma específica, en la figura 2 se presentan las taxonomías de riesgos como son las taxonomías integrales, el enfoque de riesgo social y las taxonomías regulatorias.

Figura 2
Taxonomía de riesgos con un enfoque integral



Fuente: Elaboración propia con apoyo de IA.

2.2.2 Evaluación del Impacto Ético de la IA

El uso de la inteligencia artificial generativa y de herramientas impulsadas por IA en la comunicación han generado impactos de diferente orden, los cuales se han analizado, categorizado y evaluado de la siguiente manera: Evaluaciones de impacto de IA; evaluaciones de impacto en derechos humanos; evaluaciones de impacto social y evaluaciones de impacto de igualdad (Stahl et al., 2023). Este tipo de evaluaciones tienen por alcance identificar y mitigar los riesgos asociados a la IA.

2.2.3 Vulnerabilidades y Sesgos en Sistemas de IA

Los sistemas de IA son propensos a reproducir sesgos presentes en los datos de entrenamiento. Estos sesgos pueden estar relacionados con el género, la raza, la orientación política, entre otros factores, lo que puede derivar en escenarios de discriminación, desinformación e injusticia. Además, los problemas de seguridad y las vulnerabilidades técnicas de estos sistemas pueden facilitar el uso malintencionado del conocimiento generado por la IA.

2.3 Desinformación

La desinformación es uno de los desafíos éticos de mayor análisis en el contexto del uso de la inteligencia artificial generativa en los procesos de comunicación. Este apartado estudia el alcance, la categorización y mecanismos de mitigación de la desinformación.

2.3.1 Conceptualización y Alcance

La desinformación creada por la inteligencia artificial generativa se refiere a la generación y difusión de contenido falso o engañoso con el fin de manipular a la opinión pública. Esto incluye: Texto, imágenes, video y audio, que son utilizados para alterar la percepción objetiva de la opinión pública. Las funcionalidades avanzadas de la inteligencia artificial generativa han ampliado la capacidad, sofisticación y uso de esta práctica.

2.3.2 Taxonomía de la Desinformación Generada por Inteligencia Artificial Generativa

La desinformación generada por inteligencia artificial generativa puede clasificarse en las categorías visualizadas en la Figura 3. Particularmente, en la Figura 3 se visualizan la taxonomía de la desinformación generada por IAG como es la Información falsa intencional, contenido manipulado, el contenido fuera de contexto, contenido sintético.

Figura 3
Taxonomía de la desinformación generada por IAG

Fuente: Elaboración propia apoyada con IA.

Información falsa intencional: Contenido creado con el propósito de manipular la percepción de la opinión pública.

Contenido manipulado: Modificación parcial de información verídica con el fin de distorsionar su significado.

Contenido fuera de contexto: Contenido generado sin la especificación de un contexto objetivo y preciso.

Contenido sintético: Información generada íntegramente por inteligencia artificial generativa, desde la creación de datos sintéticos hasta la producción de contenidos basados en estos

datos. Instancias de este tipo de desinformación incluyen las deepfakes y los textos automatizados.

En forma complementaria, otro enfoque a partir del cual se puede clasificar la desinformación es el de la intencionalidad, en el cual se puede clasificar en desinformación intencional (desinformation) y desinformación no intencional (misinformation). La desinformación intencional se caracteriza por tener la intención deliberada de engañar, creación y difusión intencional de información falsa, uso estratégico del conocimiento generado por IA para manipular, por definir objetivos específicos para desinformar, planificación en la distribución. Por otro lado, la desinformación no intencional se caracteriza por difusión no intencional de información incorrecta, errores o sesgos no deliberados, uso inadvertido de IA que genera inexactitudes, propagación por desconocimiento, distribución orgánica sin planificación.

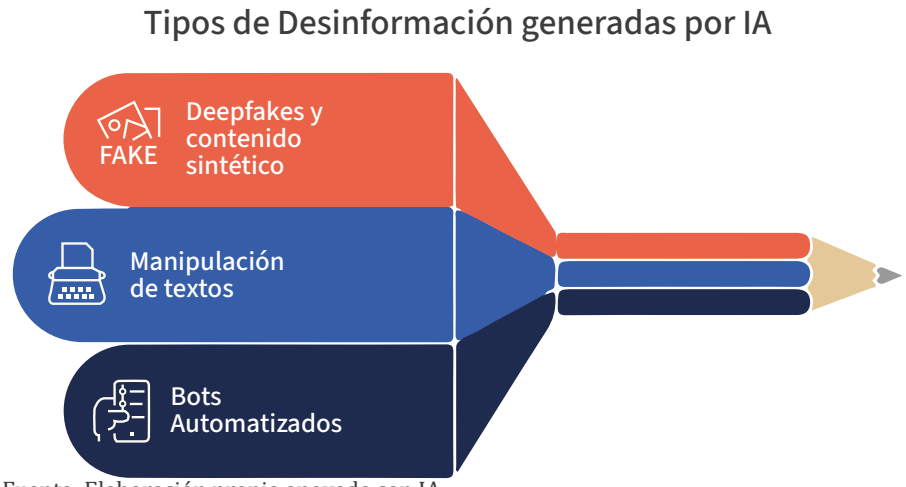
2.3.3 Mecanismos de Generación y Difusión

La inteligencia artificial facilita la generación de desinformación a través de inteligencias artificiales generativas y herramientas impulsadas por inteligencia artificial para la generación de imágenes, audios y vídeos. En forma complementaria, la difusión de este tipo de contenidos se puede hacer a través de bots.

2.4 Tipos de Desinformación Generadas por Inteligencia Artificial

La inteligencia artificial ha dado lugar a nuevos tipos de desinformación como los ilustrados en la Figura 4. En forma específica, en la figura 4 se visualizan los tipos de desinformación generados por IA como son las deepfakes y contenido sintético, la manipulación de textos y los bots automatizados.

Figura 4
Tipos de desinformación



Fuente: Elaboración propia apoyada con IA.

Deepfakes y contenido sintético: Los deepfakes son videos o audios generados por inteligencia artificial que imitan de manera convincente la apariencia y la voz de personas reales. Por otro lado, el contenido sintético hace referencia a contenido creado a partir de datos sintéticos (generados por la inteligencia artificial).

Manipulación de textos y narrativas artificiales: Información manipulada a través de inteligencias artificiales generativas, las cuales pueden crear narrativas falsas, noticias engañosas y campañas de desinformación.

Bots y difusión automatizada: Los bots son aplicaciones automatizadas que interactúan con redes sociales para difundir contenidos específicos y pueden contribuir a la distribución de desinformación.

2.5 Impacto Social de la Desinformación

La desinformación generada por la difusión del conocimiento producido por las inteligencias artificiales generativas tiene implicaciones e impactos a nivel de los procesos democráticos, el discurso público y la confianza de la sociedad.

2.5.1 Efectos en los Procesos Democráticos

La manipulación de la información generada con inteligencia artificial generativa en el contexto de los procesos democráticos puede tener implicaciones relacionadas con influenciar los procesos electorales, impactar la opinión pública y resquebrajar los principios fundamentales de la democracia, erosionando la confianza en las instituciones y figuras públicas (Ustinovich, 2024).

2.5.2 Implicaciones en el Discurso Público

La generación intencional de desinformación puede conducir a polarizar el debate público en contiendas electorales, impactar la integridad y calidad de la información disponible y fomentar la desconfianza en los medios de comunicación tradicionales (Peña-Fernández et al., 2023).

2.5.3 Impacto en la Confianza de la Sociedad

La ausencia de procesos de verificación y validación del conocimiento creado por las inteligencias artificiales generativas y la incapacidad para distinguir entre conocimiento científicamente valido y conocimiento fiable generado por IAG puede conducir potencialmente a disminuir la confianza en las instituciones, en los medios de comunicación y en las relaciones interpersonales, lo cual genera escenarios de incertidumbre en los diferentes sectores de la sociedad.

2.6 Estrategias de Mitigación y Control de la desinformación

Ante el incremento del uso de la desinformación en diferentes sectores de la sociedad, es necesario implementar medidas de mitigación con el fin de controlar y prevenir los riesgos asociados a la desinformación.

2.6.1 Herramientas de Detección y Verificación

La investigación y desarrollo en tecnologías de detección como las herramientas de análisis forense para detección de deepfakes y

sistemas de verificación de hechos se han convertido en estrategias de mitigación de la desinformación.

2.6.2 Alfabetización en Inteligencia Artificial

La alfabetización en inteligencia artificial es esencial para capacitar a los usuarios en los procesos de identificación de contenido falso y estrategias de verificación y validación del conocimiento generado con el fin de minimizar el riesgo de las alucinaciones y aumentar la precisión de las respuestas de los sistemas de IAG.

2.6.3 Marcos de Gobernanza

Las organizaciones internacionales y los gobiernos nacionales deben diseñar políticas que regulen y controlen el uso de la inteligencia artificial en los procesos de comunicación, fomentando el uso responsable y ético del conocimiento generado por la inteligencia artificial generativa.

2.7 Perspectivas del Uso de la Inteligencia Artificial Generativa en los Procesos de Comunicación

Esta sección concluye con una reflexión acerca de las tendencias emergentes, los desafíos anticipados y las recomendaciones para abordar los aspectos éticos del uso del conocimiento generado por la inteligencia artificial en el futuro.

2.7.1 Tendencias Emergentes

La rápida evolución en el diseño e implementación de sistemas de inteligencia artificial generativa ha mejorado significativamente sus funcionalidades, así como la precisión y calidad del conocimiento que generan. Además, el desarrollo de aplicaciones impulsadas por inteligencia artificial ha optimizado la eficiencia de los procesos en diversos sectores. No obstante, estas tecnologías también presentan riesgos más sofisticados y difíciles de identificar. Un ejemplo de ello es la percepción del conocimiento generado por la inteligencia artificial generativa como altamente fiable por parte del usuario final, lo que puede llevar a la omisión de procesos de verificación

y validación. Esta falta de control puede derivar en problemas de desinformación al momento de difundir dicho conocimiento.

2.7.2 Desafíos

Entre los principales desafíos asociados al diseño, implementación, uso y difusión del conocimiento generado por los sistemas de inteligencia artificial generativa y las herramientas impulsadas por esta tecnología, se encuentra la formulación e implementación de políticas de uso de la inteligencia artificial a nivel gubernamental. Estas políticas deben mantenerse vigentes y evolucionar en un ámbito de conocimiento que, al ser transversal a toda la sociedad, está en constante transformación.

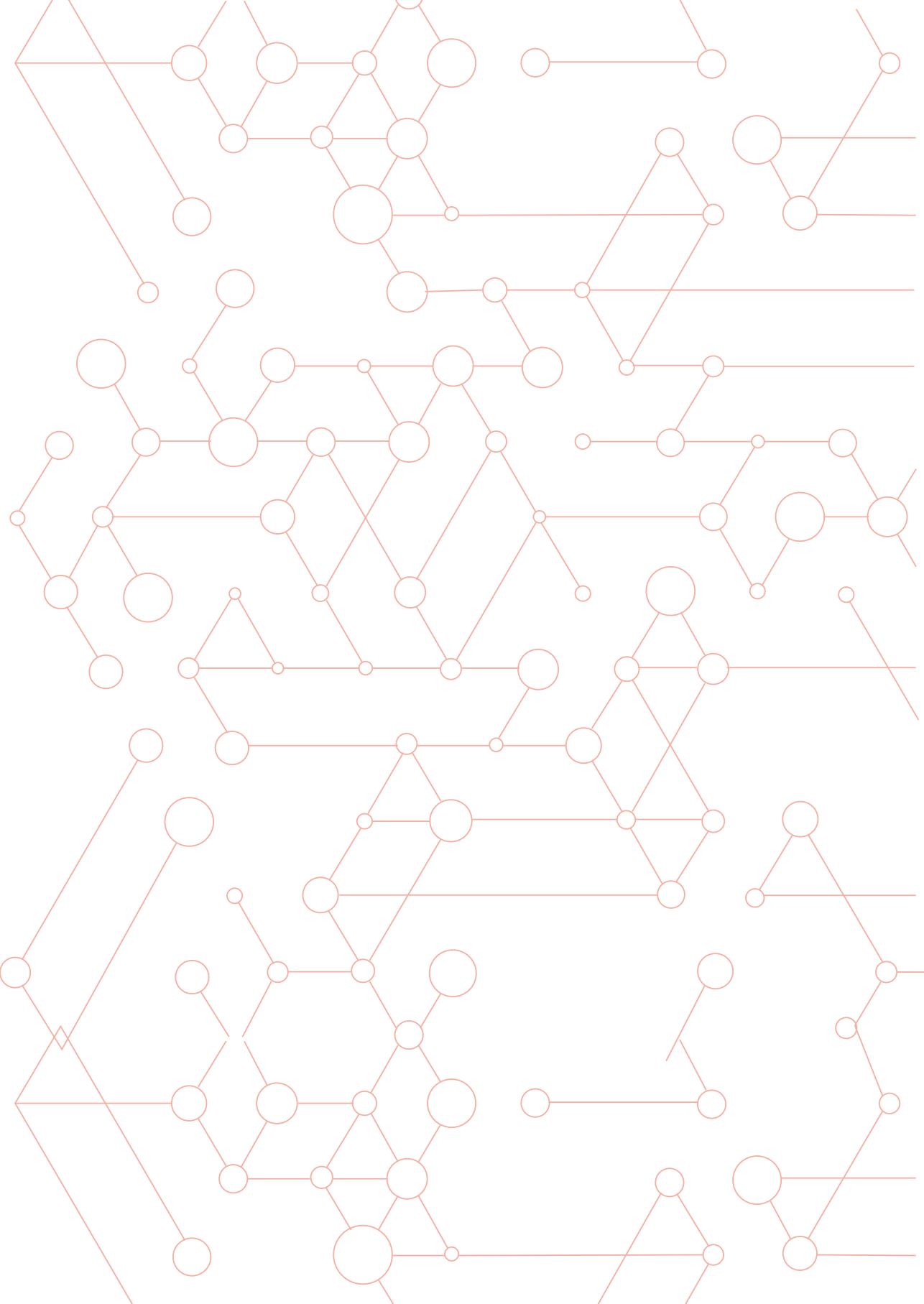
Asimismo, es fundamental diseñar, aplicar y monitorear estas políticas en el contexto de la educación superior, con un enfoque en la mitigación de la desinformación en actividades relacionadas con la comunicación y difusión del conocimiento generado por la inteligencia artificial. Por otro lado, los sesgos algorítmicos derivados del preentrenamiento de los datos y del diseño implementado por los desarrolladores de los sistemas de inteligencia artificial generativa representan otro desafío clave que requiere análisis y estudio en los próximos meses. Finalmente, la protección de los derechos humanos frente al uso indebido del conocimiento creado por estas tecnologías constituye un reto adicional que debe ser considerado con especial atención.

2.7.3 Recomendaciones Éticas

En este apartado, es fundamental que las investigaciones sobre aspectos éticos se enfoquen en temas relacionados con los principios de equidad, transparencia, responsabilidad, beneficencia y gestión de la desinformación derivada del uso del conocimiento generado por los sistemas de inteligencia artificial generativa. Asimismo, es recomendable seguir fortaleciendo las investigaciones en ética de la inteligencia artificial y continuar con la mejora continua de las políticas de regulación de su uso a nivel internacional, considerando su impacto transversal y multicultural.

CAPÍTULO III

ACERCAMIENTO METODOLÓGICO A TRAVÉS DE LA REVISIÓN SISTEMÁTICA DE LITERATURA SOBRE RIESGOS DEL USO DE LA INTELIGENCIA ARTIFICIAL EN LA COMUNICACIÓN E INVESTIGACIÓN CIENTÍFICA

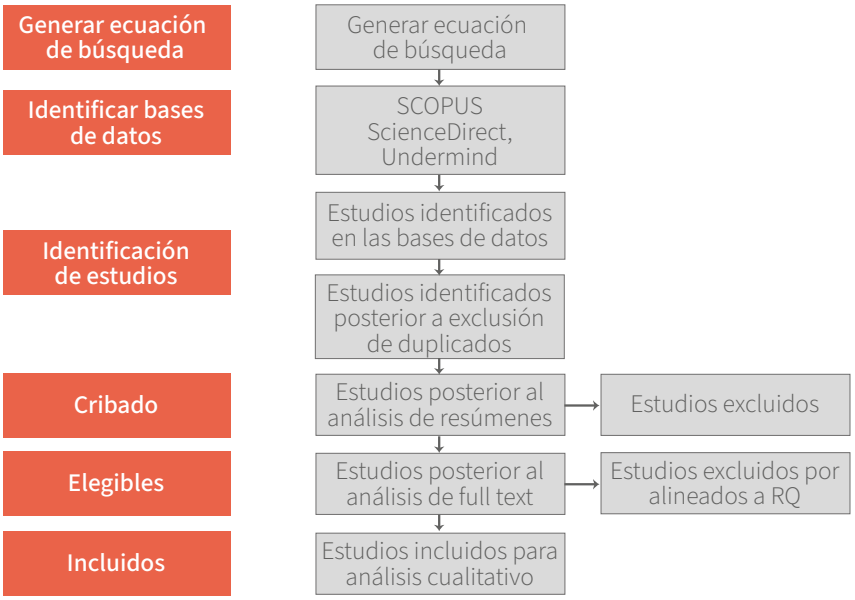


CAPÍTULO III

ACERCAMIENTO METODOLÓGICO A TRAVÉS DE LA REVISIÓN SISTEMÁTICA DE LITERATURA SOBRE RIESGOS DEL USO DE LA INTELIGENCIA ARTIFICIAL EN LA COMUNICACIÓN E INVESTIGACIÓN CIENTÍFICA

La estrategia de investigación empleada, es una revisión sistemática de literatura la cual permitió identificar los alcances esenciales de los artículos de investigación acerca de los riesgos del uso de la inteligencia artificial en los procesos de comunicación científica, focalizando el estudio en los riesgos relacionados con el dominio de desinformación (Slattery et al., 2024). En forma específica, se orientó el proceso de revisión sistemática de literatura a través del protocolo PRISMA, como se visualiza en la Figura 5. En forma específica, se visualizan las etapas del protocolo PRISMA con una adaptación para la era de la IA.

Figura 5
Protocolo PRISMA adaptado para revisión sistemática



Fuente: Elaboración propia.

En esta investigación, se examinó la siguiente pregunta de investigación:

P1: ¿Cuáles son los subdominios del dominio de desinformación que se generan en la comunicación e investigación científica por el uso de la inteligencia artificial?

Para dar respuesta a la pregunta de investigación, se inició con la especificación de la ecuación de búsqueda.

3.1 Generación de la Ecuación de Búsqueda

Se generó la ecuación de búsqueda utilizando la inteligencia artificial generativa chatgpt por medio de un prompt, en el cual se le pide que genere una ecuación de búsqueda optimizada acerca del objeto de estudio y luego se sometió al proceso de verificación y validación por parte de los investigadores como se indica en la Tabla 1.

Tabla 1
Ecuación de búsqueda de acuerdo al objeto de estudio

Objeto de estudio	Ecuación de búsqueda
Riesgos de desinformación del uso de la inteligencia artificial en la comunicación científica	(“artificial intelligence” OR AI) AND (“misinformation” OR “disinformation”) AND (“Scientific communication”)

Fuente: Elaboración propia

3.2 Identificación de las Bases de Datos y Motores de Búsqueda

Posteriormente, se identificaron las bases de datos relevantes para el campo de las Ciencias sociales, Ciencias de la Computación y la Ingeniería, se tuvieron en cuenta aquellas que cubren las conferencias y revistas más importantes en el campo de la tecnología educativa. Adicionalmente, como criterio de selección se utilizó aquellas que se consideran referentes en el campo de la educación y la investigación siendo estas, Scopus, ScienceDirect y el motor de búsqueda Undermind.

3.3 Identificación de Artículos Iniciales en Bases de Datos y Motores de Búsqueda

Teniendo en cuenta la ecuación de búsqueda se realizó una búsqueda en cada una de las bases de datos electrónicas y en el motor de búsqueda, luego se aplicaron filtros para los años 2020-2024, las cuales arrojaron la siguiente cantidad de documentos representados en la Tabla 2.

Tabla 2
Cantidad de documentos extraídos de las bases de datos de acuerdo a la ecuación de búsqueda

Ecuación de búsqueda	SCOPUS	SCIENCEDIRECT	UNDERMIND
("artificial intelligence" OR AI) AND ("misinformation" OR "disinformation") AND ("Scientific communication")	6	25	44

Fuente: Elaboración propia

3.4 Criterios de Inclusión y Exclusión

Se implementaron criterios de inclusión (I) para formalizar la inclusión de un artículo en la revisión sistemática y de exclusión (E) para eliminar artículos del proceso de revisión sistemática.

I1: Artículos que hayan sido publicados en el dominio de tiempo del 2020-2024.

I2: Artículos cuyo campo de estudio fuesen las Ciencias sociales, las Ciencias computacionales o la ingeniería.

I3: Artículos cuya tipología fuese artículo de revisión o artículo de investigación.

En cuanto a los criterios de exclusión se identificaron los siguientes:

E1: Artículos en los cuales al revisar el título y el abstract no tenían un aporte significativo al objeto de estudio.

E2: Artículos en los que al revisar de forma completa el trabajo se identifica que no está completamente alineado al objeto de estudio y a la pregunta de investigación.

3.5 Detección de Duplicados

Una vez se obtuvieron los resultados de la aplicación de la ecuación de búsqueda en las bases de datos SCOPUS, ScienceDirect y Undermind se procedió a mezclar los ítems que se alcanzaron como resultado de las búsquedas específicas en cada base de datos, a continuación, se identificaron los artículos duplicados (estos aparecen como resultado en la búsqueda en las bases de datos). En esta etapa de revisión sistemática, como resultado se obtuvo que se identificaron dos duplicados procediendo a eliminar uno de ellos, con lo cual el universo de artículos a revisar es de 74.

3.6 Proceso de Screening a Nivel de Título y Abstract

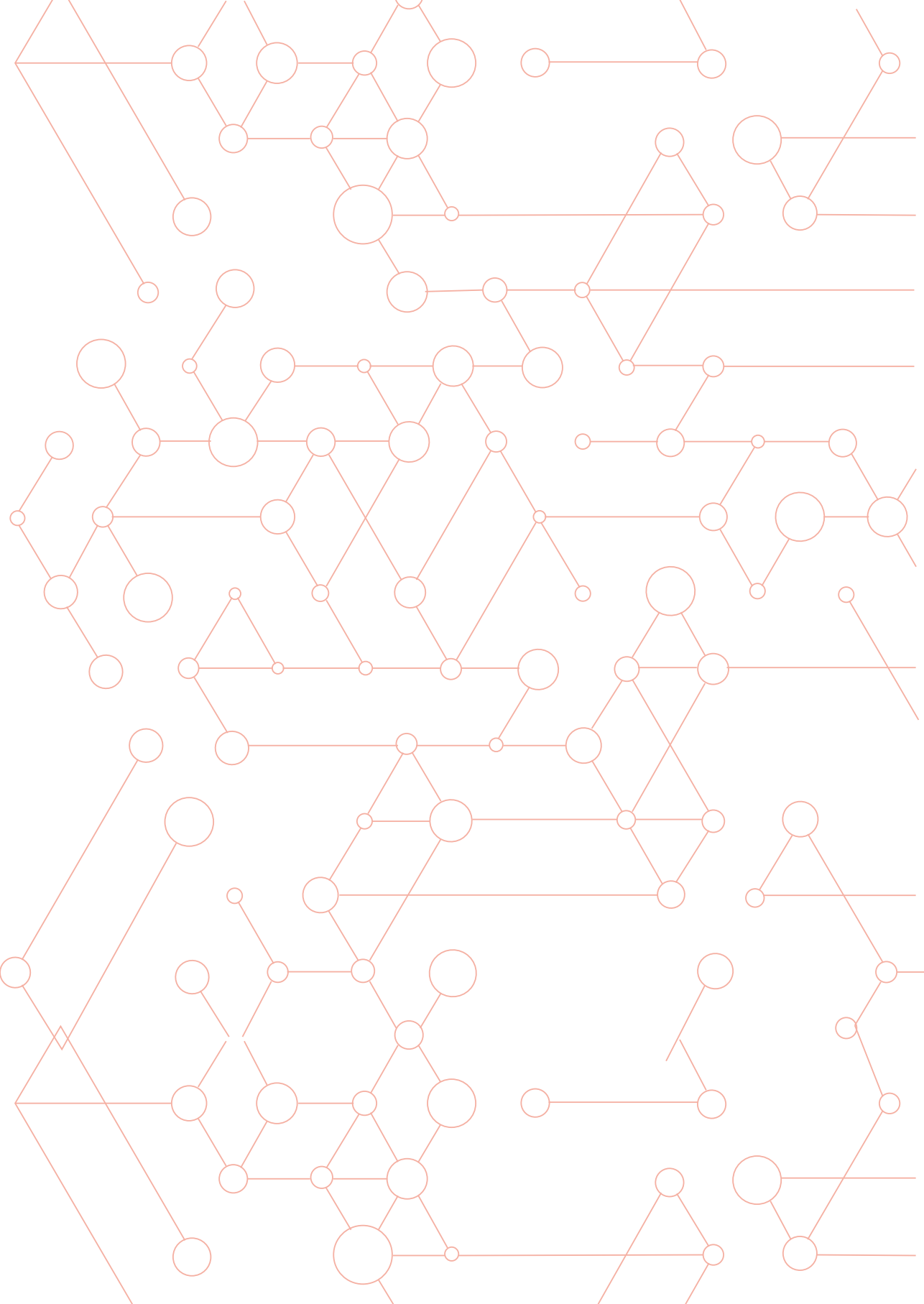
El proceso de Screening a nivel de título y abstract consiste en revisar el universo de artículos en un primer nivel, utilizando como criterios de revisión la alineación del título y el abstract con el objeto de estudio.

3.7 Proceso de Screening a Texto Completo

El proceso de screening a texto completo consiste en revisar de forma integral los artículos que quedaron después de analizar los ítems de resultados aplicando el criterio de exclusión 1 (E1), es decir, los artículos que fueron seleccionados después de la etapa de screening a nivel de título y abstract.

CAPÍTULO IV

EL DOMINIO DE RIESGOS DE DESINFORMACIÓN Y SUBDOMINIOS ASOCIADOS



CAPÍTULO IV

EL DOMINIO DE RIESGOS DE DESINFORMACIÓN Y SUBDOMINIOS ASOCIADOS

4.1 Revisión de Literatura

La revisión sistemática de literatura, permitió la integración y análisis de los artículos de los resultados de la búsqueda en las bases de datos SCOPUS, ScienceDirect y el motor de búsqueda impulsado por inteligencia artificial Undermind, conformándose un universo de 75 artículos y posteriormente se llevó a cabo el proceso de detección de duplicados en el cual se eliminó un artículo duplicado, con lo cual el universo quedó estructurado por 74 artículos.

Posteriormente, se procedió a aplicar los criterios de inclusión I1, I2, I3 y el criterio de exclusión E1 en el primer nivel de revisión (Screening a nivel de título y abstract), obteniendo como resultado 25 artículos, los cuales se especifican en la Tabla 3.

En el contexto de la tabla 3, los identificadores nombrados con Rn hacen alusión a la referencia en el orden en que fueron identificadas y seleccionadas para el proceso de screening, de tal manera que R1 corresponde al identificador del primer artículo seleccionado para la revisión sistemática, R2 al segundo artículo seleccionado y así sucesivamente con todos los artículos identificados para revisión.

Tabla 3
Artículos seleccionados en el primer nivel de Screening

Identificador	Referencia	Título
R1	Slattery, P., Saeri, A., Grundy, E.A., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S. & Thompson, N., (2024). The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. ArXiv, 1. https://doi.org/10.48550/arXiv.2408.12622	The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence
R2	Penabad-Camacho, L., Penabad-Camacho, M. A., Mora-Campos, A., Cerdas-Vega, G., Morales-López, Y., Ulate-Segura, M., A. Méndez-Solano, Nova-Bustos, N., Vega-Solano, M. F. y Castro-Solano, M. M. Trans., (2024) Declaración de Heredia: Principios sobre el uso de la Inteligencia Artificial en la publicación científica. Revista Electrónica Educare, 28(S), 1-10. https://doi.org/10.15359/ree.28-S.19967	Heredia Declaration: Principles on the use of Artificial Intelligence in scientific publishing
R3	Carobene, A., Padoan, A., Cabitza, F., Banfi, G. & Plebani, M., (2023). Rising adoption of artificial intelligence in scientific publishing: evaluating the role, risks, and ethical implications in paper drafting and review process. Clinical chemistry and laboratory medicine, 62(5), 835–843. https://doi.org/10.1515/cclm-2023-1136	Rising adoption of artificial intelligence in scientific publishing: evaluating the role, risks, and ethical implications in paper drafting and review process
R4	Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X., (2023). AI vs. Human -- Differentiation Analysis of Scientific Content Generation. ArXiv. https://doi.org/10.48550/arXiv.2301.10416	AI vs. Human - Differentiation Analysis of Scientific Content Generation
R5	Ballester, V. G. & Vicente, C. T., (2024). El rol de la inteligencia artificial en la publicación científica: perspectivas desde la farmacia hospitalaria. Farmacia hospitalaria: órgano oficial de expresión científica de la Sociedad Española de Farmacia Hospitalaria, 48(5), 246-251. https://doi.org/10.1016/j.farma.2024.06.002	El rol de la inteligencia artificial en la publicación científica: perspectivas desde la farmacia hospitalaria

R6	Hryciw, B.N., Seely, A.J. & Kyeremanteng, K., (2023). Guiding principles and proposed classification system for the responsible adoption of artificial intelligence in scientific writing in medicine. Frontiers in Artificial Intelligence, 6, 1-5. https://doi.org/10.3389/frai.2023.1283353	Guiding principles and proposed classification system for the responsible adoption of artificial intelligence in scientific writing in medicine
R7	Rodrigues, F., Newell, R., Babu, G.R., Chatterjee, T., Sandhu, N.K., & Gupta, L., (2024). The social media Infodemic of health-related misinformation and technical solutions. Health Policy and Technology, 13(2), 100846. https://doi.org/10.1016/j.hlpt.2024.100846	The social media Infodemic of health-related misinformation and technical solutions
R8	Khalifa, M. & Albadawy, M., (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. Computer Methods and Programs in Biomedicine Update, 5, 100145. https://doi.org/10.1016/j.cmpbup.2024.100145	Using artificial intelligence in academic writing and research: An essential productivity tool
R9	Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V. & Isaac, W., (2023). Sociotechnical Safety Evaluation of Generative AI Systems. ArXiv. http://arxiv.org/abs/2310.11986	Sociotechnical Safety Evaluation of Generative AI Systems
R10	Cui, T., Wang, Y., Fu, C., Xiao, Y., Li, S., Deng, X., Liu, Y., Zhang, Q., Qiu, Z., Li, P., Tan, Z., Xiong, J., Kong, X., Wen, Z., Xu, K. & Li, Q., (2024). Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems. ArXiv. http://arxiv.org/abs/2401.05778	Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems
R11	Cunha, P.R. & Estima, J., (2023). Navigating the Landscape of AI Ethics and Responsibility. In N. Moniz, Z. Vale, J. Cascalho, C. Silva, R. Sebastião (eds) Progress in Artificial Intelligence. EPIA 2023. Lecture Notes in Computer Science (vol. 14115, pp. 92-105). Springer. https://doi.org/10.1007/978-3-031-49008-8_8	Navigating the Landscape of AI Ethics and Responsibility

R12	Deng, J., Cheng, J., Sun, H., Zhang, Z. & Huang, M., (2023). Towards Safer Generative Language Models: A Survey on Safety Risks, Evaluations, and Improvements. ArXiv, 1. http://arxiv.org/abs/2302.09270	Towards Safer Generative Language Models: A Survey on Safety Risks, Evaluations, and Improvements
R13	Hagendorff, T., (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. ArXiv. http://arxiv.org/abs/2402.08323	Mapping the Ethics of Generative AI: A Comprehensive Scoping Review
R14	Hogenhout, L., (2021). A Framework for Ethical AI at the United Nations. ArXiv. http://arxiv.org/abs/2104.12547	A framework for ethical AI at the United Nations
R15	Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N. 'mah, Gallegos, J., Smart, A., Garcia, E., & Virk, G., (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 723-741). Association for Computing Machinery. https://doi.org/10.1145/3600211.3604673	Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction.
R16	Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J, Legassick, S., Irving, G. & Gabriel, I., (2022). Taxonomy of Risks posed by Language Models. FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 214–229. https://doi.org/10.1145/3531146.3533088	Taxonomy of Risks posed by Language Models.
R17	Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A. & Gabriel, I., (2021). Ethical and social risks of harm from Language Models. ArXiv. http://arxiv.org/abs/2112.04359	Ethical and social risks of harm from language models

R18	Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V. & Isaac, W., (2023). Sociotechnical Safety Evaluation of Generative AI Systems. ArXiv. http://arxiv.org/abs/2310.11986	Sociotechnical Safety Evaluation of Generative AI Systems.
R19	Vidgen, B., Agrawal, A., Ahmed, A. M., Akinwande, V., Al-Nuaimi, N., Alfaraj, N., Alhajjar, E., Aroyo, L., Bavalatti, T., Blili-Hamelin, B., Bollacker, K., Bomassani, R., Boston, M.F., Campos, S., Chakra, K., Chen, C., Coleman, C., Coudert, Z. D., Derczynski, L., "..." Vanschoren, J., (2024). Introducing v0.5 of the AI Safety Benchmark from MLCommons. ArXiv. http://arxiv.org/abs/2404.12241	Introducing v0.5 of the AI Safety Benchmark from MLCommons
R20	Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Stevie Bergman, A., Shelby, R., Marchal, N., Griffin, C., "..." Manyika, J., (2024). The Ethics of Advanced AI Assistants. ArXiv, 1. https://doi.org/10.48550/arXiv.2404.16244	The Ethics of Advanced AI Assistants
R21	Sun, H., Zhang, Z., Deng, J., Cheng, J. & Huang, M., (2023). Safety Assessment of Chinese Large Language Models. ArXiv. https://doi.org/10.48550/arXiv.2304.10436	Safety Assessment of Chinese Large Language Models
R22	Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J. & Huang, M., (2023). SafetyBench: Evaluating the safety of Large Language Models with multiple choice questions. ArXiv. https://doi.org/10.48550/arXiv.2309.07045	SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions
R23	Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F. & Li, H., (2023). Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. ArXiv. http://arxiv.org/abs/2308.05374	Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment
R24	Electronic Privacy Information Centre., (2023). Generating Harms: Generative AI's Impact & Paths Forward. https://epic.org/documents/generating-harms-generative-ais-impact-paths-forward/	Generating Harms: Generative AI's Impact & Paths Forward

R25	Nah, F. F. H., Zheng, R., Cai, J., Siau, K. & Chen, L., (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Information Technology Case and Application Research, 25(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814	Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration
R26	Hendrycks, D. & Mazeika, M., (2022). X-Risk Analysis for AI Research. ArXiv. https://arxiv.org/abs/2206.05862	X-Risk Analysis for AI Research
R27	Infocomm Media Development Authority., (2023). Cataloguing LLM Evaluations. https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf	Cataloguing LLM Evaluations

Fuente: Elaboración propia

En el segundo nivel de revisión a texto completo se seleccionaron 24 artículos, con los cuales se procedió a realizar el proceso de análisis para dar respuesta a la pregunta de investigación como se especifica en la Tabla 4.

Tabla 4

Artículos seleccionados para revisión a texto completo del dominio de desinformación

Identificador	Especificación de riesgos del dominio de desinformación (Respuesta a pregunta de investigación)	Instancias de riesgos del dominio de desinformación (Triangulación-Análisis)
R1	Información falsa o engañosa. Contaminación del ecosistema informativo y pérdida de la realidad consensuada. Desinformación, vigilancia e influencia a gran escala.	Información falsa o engañosa: Generación de contenido incorrecto, datos de entrenamiento ruidosos, estrategias de muestreo aleatorias, bases de conocimiento desactualizadas. Contaminación del ecosistema informativo y pérdida de la realidad consensuada: Desconfianza generalizada en la información, fragmentación de la realidad compartida, amplificación de la desinformación. Desinformación, vigilancia e influencia a gran escala: Propagación de la desinformación, vigilancia masiva, manipulación de la opinión pública, censura automatizada.

R2	Sesgos inherentes. Falta de transparencia. Propagación de desinformación. Desconocimiento de los usuarios. Interacción humana.	Sesgos inherentes: Selección de datos, interpretación de resultados, diseño del modelo, feedback y ajuste. Falta de transparencia: Desconocimiento del proceso, dificultad en la reproducibilidad, confusión sobre la autoría, propagación de errores, falta de confianza. Propagación de la desinformación: Generación de contenido erróneo, amplificación de errores existentes, interpretación incorrecta de resultados, confusión en la comunicación, dependencia excesiva de la IA. Desconocimiento de los usuarios: Falta de comprensión de la tecnología, uso inapropiado de la tecnología, desconfianza de los resultados, dificultades en la evaluación de la calidad, propagación de sesgos, inadecuada protección de datos. Interacción humana: Malentendidos en la comunicación, dependencia excesiva de la IA, sesgos en la interpretación, falta de responsabilidad, interacción superficial, confusión sobre el rol de la IA.
R3	Generación de información incorrecta. Inconsistencias en las respuestas. Dependencia de contenido no verificado. Reproducción de sesgos. Desafíos en la originalidad y plagio. Impacto en la calidad de la literatura científica.	Generación de información incorrecta: Alucinaciones, falta de comprensión contextual, errores en datos específicos, confusión de términos técnicos, inconsistencias en la información. Inconsistencia en las respuestas: variabilidad en respuestas a la misma pregunta, incoherencia en el contenido, dificultades en la interpretación de datos, falta de contexto adecuado, desviaciones en el estilo y tono. Dependencia de contenido no verificado: difusión de información errónea, falta de credibilidad, reproducción de sesgos, problemas de plagio y originalidad, interpretaciones erróneas, consecuencias legales y éticas. Reproducción de sesgos: refuerzo de estereotipos, desigualdad en la representación, decisiones erróneas basados en datos sesgados, impacto en la credibilidad de la investigación, desinformación y confusión. Desafíos en la originalidad y plagio: Generación de contenido similar, dificultades en la atribución de autoría, confusión entre inspiración y copia, problemas legales de derecho de autor, desafíos en la evaluación de la originalidad, impacto en la integridad académica.

		Impacto en la calidad de la literatura científica: disminución en la rigurosidad científica, variabilidad en la calidad del contenido, propagación de errores, desafíos en la revisión de pares, impacto en la credibilidad de las publicaciones, reducción de la diversidad de perspectivas, desafíos éticos y de integridad.
R4	Errores Factuales. Inconsistencias y controversias. Manipulación de la opinión pública. Contaminación de modelos de lenguaje. Falta de profundidad y perspectiva. Desafíos en la integridad de la publicación científica.	Errores factuales: difusión de información incorrecta, desconfianza en la literatura científica, impacto en la toma de decisiones, contaminación de bases de datos, desinformación en medios de comunicación. Inconsistencias y controversias: Generación de contenido contradictorio, confusión en la interpretación de resultados, controversias en la comunidad científica, desconfianza en la autoridad de la fuente, dificultades en la revisión por pares. Manipulación de la opinión pública: Generación de contenido sesgado, desinformación y fake news, amplificación de narrativas extremas, manipulación de debates públicos, impacto en la confianza en los medios. Contaminación de modelos de lenguaje: incorporación de información errónea, reforzamiento de sesgos, generación de contenido inconsistente, dificultades en la toma de decisiones, desconfianza en la tecnología, impacto en la calidad de la investigación, Falta de profundidad y perspectiva: Superficialidad en el análisis, falta de contexto, pérdida de perspectivas diversas, desconexión de realidades prácticas, reducción de la creatividad, desinformación por simplificación. Desafíos en la integridad de la publicación científica: plagio y autenticidad, falsificación de resultados, revisión por pares inadecuada, desinformación y confusión, impacto en la reputación de revistas, dificultades en la citación y referenciación.
R5	Facilidad para el Plagio. Incorporación de sesgos no detectados. Generación de información falsa. Dificultad para verificar la veracidad.	Facilidad para el plagio: Reproducción de texto existente, parafraseo inadecuado, falta de citación adecuada, dependencia de contenido generado por IA. Incorporación de sesgos no detectados: sesgos en los datos de entrenamiento, refuerzo de estereotipos, desigualdad en la representación, falta de contexto cultural. Generación de información falsa: Alucinaciones de IA, falta de verificación de hechos, confusión entre hechos y opiniones, desinformación intencionada, impacto en la credibilidad científica, dificultad para rectificar errores.

		Dificultad para verificar la veracidad: Generación de contenido convincente, falta de fuentes claras, ambigüedad en la información, inconsistencias en los datos, dependencia de datos de entrenamiento, desafíos en la detección de plagio.
R6	Sesgos en los datos de entrenamiento. Inexactitudes en el contenido generado. Dependencia excesiva de la IA. Falta de transparencia. Desinformación intencionada.	Sesgos en los datos de entrenamiento: Subrepresentación de grupos, refuerzo de estereotipos, falta de contexto cultural. Inexactitudes en el contenido generado: interpretación errónea de datos, falta de actualización, generalización inapropiada, errores en la síntesis de información, ambigüedad en el lenguaje. Dependencia excesiva de la IA: desarrollo de habilidades críticas, reducción de la creatividad, deshumanización del proceso de escritura, falta de contexto y comprensión. Falta de transparencia: confusión sobre la autoría, dificultad para evaluar la calidad, desconfianza en la investigación, sesgos no reconocidos, falta de responsabilidad. Desinformación intencionada: Propagación de información errónea, erosión de la confianza en la ciencia, impacto en la salud pública, manipulación de la opinión pública, desigualdades en el acceso a la información.
R7	Diseminación de información errónea. Falta de verificación de fuentes. Polarización de la información. Manipulación emocional. Dificultad para distinguir entre fuentes legítimas y generadas por IA. Impacto en la confianza pública.	Diseminación de información errónea: creación de contenido engañoso, reforzamiento de mitos y rumores, difusión rápida en redes sociales, falta de contexto. Falta de verificación de fuentes: propagación de información falsa, confusión y desinformación, credibilidad comprometida, manipulación de la opinión pública. Polarización de la información: creación de burbujas informativas, aumento de la desconfianza, extremismo en opiniones, desinformación dirigida. Manipulación emocional: difusión de desinformación, reacción desproporcionada, polarización de opiniones, desconfianza en fuentes legítimas, comportamientos de riesgo. Dificultad para distinguir entre fuentes legítimas y generadas por IA: Propagación de desinformación, confusión y desconfianza, manipulación de la opinión pública, dificultad en la verificación de hechos. Impacto en la confianza pública: desconfianza en las instituciones, polarización social, rechazo de información basada en evidencia, aumento de teorías de conspiración.

R8	Generación de contenido erróneo. Plagio y fabricación. Falta de contexto. Dependencia de la tecnología. Desinformación intencionada.	Generación de contenido erróneo: Inexactitudes factuales, desactualización de información, interpretaciones erróneas. Plagio y fabricación: Reproducción de texto plagiado, fabricación de datos. Falta de contexto: Interpretaciones erróneas, omisión de información crítica, generalización inapropiada. Dependencia de la tecnología: Reducción de habilidades críticas, desconexión de la creatividad humana, vulnerabilidad a errores de la IA, falta de comprensión del proceso. Desinformación intencionada: Creación de contenido falso, manipulación de datos, propagación de rumores, sesgo intencionado.
R9	Generación de contenido engañoso. Manipulación de la opinión pública. Sesgos en los datos. Dificultad para verificar la autenticidad. Impacto en la toma de decisiones.	Generación de contenido engañoso: Desinformación viral, creación de falsedades convincentes, manipulación de datos científicos. Manipulación de la opinión pública: Creación de narrativas falsas, amplificación de desinformación, desviación de la atención, polarización social. Sesgos en los datos: Reproducción de estereotipos, desigualdad en la representación, desinformación basada en datos sesgados, impacto en la toma de decisiones. Dificultad para verificar la autenticidad: Confusión entre contenido real y falso, desafíos en la verificación de fuentes, aumento de la desconfianza, manipulación de la opinión pública. Impacto en la toma de decisiones: Decisiones basadas en información errónea, falta de transparencia, dependencia de la IA, desigualdades en la toma de decisiones, impacto en la responsabilidad.
R10	Factualidad y sesgo. Manipulación de la información. Responsabilidad y ética. Detección de desinformación. Impacto en la percepción pública.	Factualidad y sesgo: Generación de información incorrecta, sesgo en los datos de entrenamiento, interpretación errónea de la información, refuerzo de estereotipos. Manipulación de la información: Creación de noticias falsas, desinformación intencionada, alteración de contextos, sesgo en la representación de hechos, manipulación de imágenes y videos. Responsabilidad y ética: Falta de transparencia, responsabilidad por el contenido generado, uso indebido de la tecnología, impacto en la privacidad.

		Detección de desinformación: Falsos positivos, falsos negativos, sesgo en los algoritmos, manipulación de la detección, dependencia de fuentes externas. Impacto en la percepción pública: Desconfianza en la información, polarización social, manipulación de la opinión pública, reacciones emocionales, impacto en la salud pública.
R12	Generación de contenido falso. Manipulación de la información. Falta de transparencia. Reforzamiento de sesgos. Desinformación intencionada. Impacto en la credibilidad científica.	Generación de contenido falso: Alucinaciones, desinformación en temas críticos, alteración de resultados científicos, propagación de teorías de conspiración. Manipulación de la información: Sesgo en la presentación de datos, alteración de resultados de investigación, desinformación selectiva, propagación de información errónea, creación de narrativas engañosas, descontextualización de citas y fuentes. Falta de transparencia: Opacidad en los algoritmos, dificultad para rastraer fuentes, sesgo inadvertido, responsabilidad difusa. Reforzamiento de sesgos: Reproducción de estereotipos, amplificación de desigualdades, desinformación basada en sesgos, falta de representación, decisiones sesgadas en aplicaciones prácticas. Desinformación intencionada: Creación de contenido falso, manipulación de opiniones, propagación de teorías de conspiración, desviación de la atención pública. Impacto en la credibilidad científica: Difusión de información errónea, desconfianza en la ciencia, hallazgos no verificados, polarización de opiniones, impacto en políticas públicas, dificultades en la comunicación científica.
R13	Generación de contenido falso. Manipulación de la información. Dificultad en la verificación. Impacto en la credibilidad. Efectos en la salud pública.	Generación de contenido falso: Alucinaciones, referencias fabricadas, desinformación intencionada, propagación de mitos. Manipulación de la información: Sesgo en la generación de contenido, alteración de datos científicos, descontextualización de información, creación de narrativas engañosas. Dificultad en la verificación: Autenticidad en las fuentes, confusión entre contenido real y contenido falso, falta de rastreabilidad. Impacto en la credibilidad: Desconfianza generalizada, erosión de la autoridad de expertos, dificultad para establecer hechos, impacto en la comunicación científica.

		Efectos en la salud pública: Desinformación sobre tratamientos médicos, propagación de mitos y rumores, consecuencias de consejos de salud inadecuados.
R14	Erosión de la verdad compartida. Falta de transparencia. Creación de contenido falso. Manipulación de la información. Desigualdad en el acceso a la información.	Erosión de la verdad compartida: Burbujas de filtro, desinformación viral, manipulación de contenido. Falta de transparencia: Decisiones de caja negra, dificultad para auditar, sesgo inadvertido. Creación de contenido falso: Deep fakes, artículos y noticias falsas, manipulación de imágenes, propaganda y desinformación. Manipulación de la información: Desinformación coordinada, alteración de narrativas, sesgo algorítmico. Desigualdad en el acceso a la información: Brecha digital, concentración de poder, desigualdad en la alfabetización digital.
R15	Generación de contenido erróneo. Falta de contexto. Sesgos en los datos. Desconfianza en la información. Dificultad para verificar fuentes.	Generación de contenido erróneo: Inexactitudes factuales, interpretaciones erróneas de datos, descontextualización, confusión de terminología. Falta de contexto: Interpretaciones erróneas, generalizaciones inapropiadas, desconexión de relevancia, omisión de información crítica. Sesgo en los datos: Reproducción de estereotipos, desigualdad en la representación, resultados inexactos, limitaciones en la diversidad de perspectivas. Desconfianza en la información: Erosión de la credibilidad, resistencia a la innovación, propagación de desinformación. Dificultad para verificar fuentes: Propagación de información errónea, desconfianza generalizada, desinformación y fake news, dificultades en la toma de decisiones.
R16	Generación de información errónea. Confusión entre hechos y opiniones. Amplificación de sesgos. Falta de transparencia. Desconfianza en fuentes autorizadas.	Generación de información errónea: Inexactitud factual, descontextualización, confusión de datos, reproducción de mitos o creencias erróneas. Confusión entre hechos y opiniones: Presentación de opiniones como hechos, ambigüedad en el lenguaje, falta de contexto, reforzamiento de sesgos. Amplificación de sesgos: Reproducción de estereotipos, desigualdad en el rendimiento, refuerzo de normas sociales dañinas, sesgo en la toma de decisiones, desinformación y polarización, dificultad para identificar sesgos.

	Desinformación intencionada.	Falta de transparencia: Dificultad para comprender el funcionamiento del modelo, desconfianza en la tecnología, imposibilidad de auditar resultados, responsabilidad difusa, dificultad para identificar sesgos.
R17	Predicciones incorrectas. Efectos en la confianza. Dificultad para distinguir la verdad. Manipulación intencionada. Impacto en la percepción pública.	Desconfianza en fuentes autorizadas: Erosión de la credibilidad, desinformación generalizada, polarización de opiniones, dificultad para tomar decisiones informadas, aumento de la vulnerabilidad a la manipulación.
R19	Generación de contenido falso. Manipulación de la opinión pública. Dificultad para verificar fuentes. Sesgos y estereotipos. Impacto en la toma de decisiones. Riesgos de seguridad.	Desinformación intencionada: Manipulación de la opinión pública, polarización social, desconfianza en instituciones, impacto en la salud pública, desestabilización de la democracia.
R20	Generación de contenido falso. Manipulación de la información. Amplificación de sesgos. Desinformación intencionada. Dificultad para verificar fuentes. Impacto en la toma de decisiones.	Generación de contenido falso: Creación de artículos científicos falsos, falsificación de resultados de investigación, suplantación de identidad de autores, desinformación en redes sociales, creación de datos estadísticos falsos. Manipulación de la información: Alteración de datos originales, descontextualización de información, sesgo en la presentación de resultados, creación de narrativas engañosas, falsificación de citas y fuentes. Amplificación de sesgos: Reforzamiento de estereotipos, desigualdad en la representación, discriminación en la toma de decisiones, generación de contenido tóxico, sesgo en la información científica.

		Desinformación intencionada: Creación de contenido falso, manipulación de imágenes y videos, difusión de rumores, campañas de desinformación coordinadas, generación de contenido sensacionalista, falsificación de fuentes, descontextualización de información. Dificultad para verificar fuentes: Propagación de información errónea, desconfianza en los medios, dificultades en la toma de decisiones, manipulación de la opinión pública, confusión y desinformación, aumento de la polarización. Impacto en la toma de decisiones: Decisiones basadas en informaciones erróneas, falta de contexto, dependencia de la IA, sesgo en la toma de decisiones, desinformación y manipulación, incertidumbre en resultados.
R21	Generación de contenido erróneo. Sesgos en la información. Falta de contexto. Manipulación intencional. Dificultad para verificar Fuentes. Impacto en la credibilidad.	Generación de contenido erróneo: Inexactitud de datos, interpretaciones erróneas, confusión de conceptos, generación de citas falsas, generalización inapropiada. Sesgos en la información: Refuerzo de estereotipos, desigualdad en la representación, preferencia por perspectivas dominantes, sesgo de confirmación, interpretaciones culturales. Falta de contexto: Malentendidos, inexactitud en la información, desinformación, interpretaciones sesgadas, dificultad en la toma de decisiones, confusión en la comunicación. Manipulación intencional: Desinformación, propaganda, sesgo en la información, creación de confusión, explotación de vulnerabilidades, desestabilización social, alteración de la opinión pública, fomento de la polarización. Dificultad para verificar fuentes: Desinformación, confusión del usuario, propagación de rumores, impacto en la toma de decisiones, erosión de la credibilidad, manipulación de la opinión pública. Impacto en la credibilidad: Desconfianza generalizada, erosión de la autoridad de fuentes, propagación de información errónea, impacto en la toma de decisiones, desconfianza en la tecnología, dificultades en la comunicación, polarización social.

R22	Generación de contenido erróneo. Sesgo en los datos. Falta de contexto. Manipulación intencional. Dificultad en la verificación. Impacto en la percepción pública.	Generación de contenido erróneo: Inexactitudes factuales, desactualización de información, interpretaciones erróneas, generalizaciones inapropiadas, confusión de terminología, falta de citación adecuada. Sesgo en los datos: Representación desigual, sesgo de confirmación, falta de diversidad en las fuentes, interpretación sesgada de resultados, generalización de estereotipos. Falta de contexto: Interpretaciones erróneas, ambigüedad en el lenguaje, desconexión cultural, inadecuación de respuestas, pérdida de matices. Manipulación intencional: Desinformación, sesgo intencional, generación de contenido malicioso, Phishing y fraude, manipulación de opiniones, alteración de resultados. Dificultad en la verificación: Falta de transparencia, inconsistencia en resultados, dificultad para rastrear fuentes, ambigüedad en la información, desactualización de datos. Impacto en la percepción pública: Desinformación, sesgo en la información, manipulación de opiniones, estigmatización, confianza en la tecnología, reacciones emocionales, polarización social.
R23	Propagación de información errónea. Falta de verificación de hechos. Manipulación de la percepción pública. Sesgo en la información. Dificultad para discernir fuentes confiables. Desconfianza en la comunicación científica.	Propagación de información errónea: Generación de contenido falso, descontextualización de información, reforzamiento de mitos y desinformación existente, difusión viral en redes sociales. Falta de verificación de hechos: Creación de narrativas sesgadas, desinformación estratégica, amplificación de teorías de conspiración, desviación de la atención. Manipulación de la percepción pública: Generación de contenido engañoso, propagación de estereotipos, desinformación dirigida, alteración de la narrativa pública. Sesgo en la información: Representación desproporcionada, amplificación de estereotipos, desinformación y noticias falsas, sesgo en la selección de datos, falta de contexto, enfoques en temas sensacionalistas. Dificultad para discernir fuentes confiables: Desinformación, falta de verificación de hechos, confusión entre fuentes legítimas y no legítimas, manipulación de contenido, sesgo en la representación de información, dificultad para evaluar la autoridad de las fuentes.

		Desconfianza en la comunicación científica: Difusión de información errónea, sensacionalismo, falta de contexto, confusión entre ciencia y pseudociencia, manipulación de datos, desigualdad en la representación de la ciencia, efecto de burbuja informativa, desconfianza en las instituciones científicas.
R25	Alucinaciones de la IA. Falta de verificación. Sesgos en los datos. Confusión en la autoría. Desinformación intencionada.	Alucinaciones de la IA: Generación de información ficticia, inexactitudes en la literatura científica, consecuencias en la toma de decisiones. Falta de verificación: Difusión de información no validada, confusión sobre la calidad de la información, impacto en la credibilidad de la información. Sesgos en los datos: Reproducción de estereotipos, desigualdad en la representación, desinformación en contextos sensibles. Confusión en la autoría: Plagio y falta de atribución, desdibujamiento de la responsabilidad, confusión en la propiedad intelectual. Desinformación intencionada: Creación de contenido engañoso, manipulación de la opinión pública, erosión de la confianza en la información.
R26	Amplificación de la desinformación. Dificultad en la verificación. Manipulación intencional. Impacto en la credibilidad. Responsabilidad ética.	Amplificación de la desinformación: Creación de contenido falso, personalización de desinformación, escalabilidad de desinformación, manipulación de narrativas. Dificultad de verificación: Confusión entre contenido real y falso, falta de fuentes confiables, desinformación viral. Manipulación intencional: Creación de fake news, campañas de desinformación, suplantación de identidad, alteración de contenido, desviación de narrativas. Impacto en la credibilidad: Desconfianza generalizada, erosión de la autoridad, dificultad para verificar fuentes, polarización de opiniones. Responsabilidad ética: Toma de decisiones automatizada, sesgo algorítmico, falta de responsabilidad, privacidad y vigilancia, manipulación de comportamientos, desigualdad en el acceso a la tecnología.

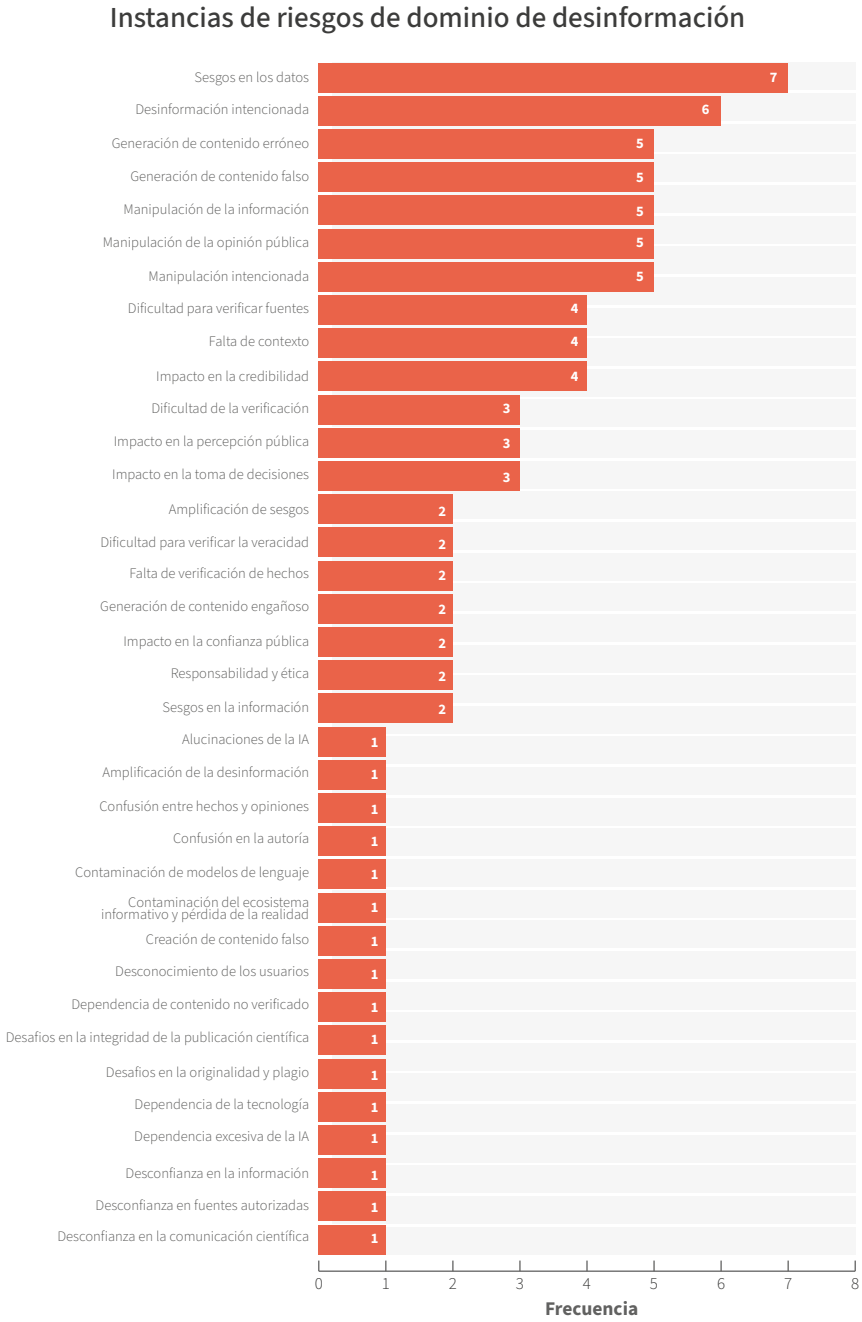
R27	Exactitud de la información. Sesgo en los datos. Falta de transparencia. Manipulación intencionada. Descontextua-lización. Impacto en la confianza pública.	Exactitud de la información: Generación de información incorrecta, confusión entre hechos y opiniones, desactualización de la información, interpretación errónea de datos. Sesgo en los datos: Representación inadecuada, refuerzo de estereotipos, falta de diversidad en los datos, desigualdad en la información. Falta de transparencia: Opacidad en los procesos de decisión, dificultad para verificar la información, incapacidad para identificar sesgos, desconocimiento de las limitaciones del modelo. Manipulación intencionada: Desinformación, propaganda, creación de contenido engañoso, manipulación de resultados de búsqueda, suplantación de identidad. Descontextualización: Interpretación errónea de la información, pérdida de matices culturales, generación de contenido ambiguo, manipulación de narrativas, desinformación. Impacto en la confianza pública: Desconfianza en la información, erosión de la credibilidad de las fuentes, manipulación de opiniones, desinformación y fake news, impacto en la participación cívica.

Fuente: Elaboración propia

Como producto del proceso de revisión sistemática de literatura se obtuvo una lista de 73 instancias de riesgos del dominio de desinformación con su respectiva frecuencia de aparición en los documentos con los cuales se llevó a cabo el proceso de revisión, inicialmente se visualizan los riesgos del dominio de desinformación con frecuencia más alta como se muestra en la Figura 6.

Figura 6

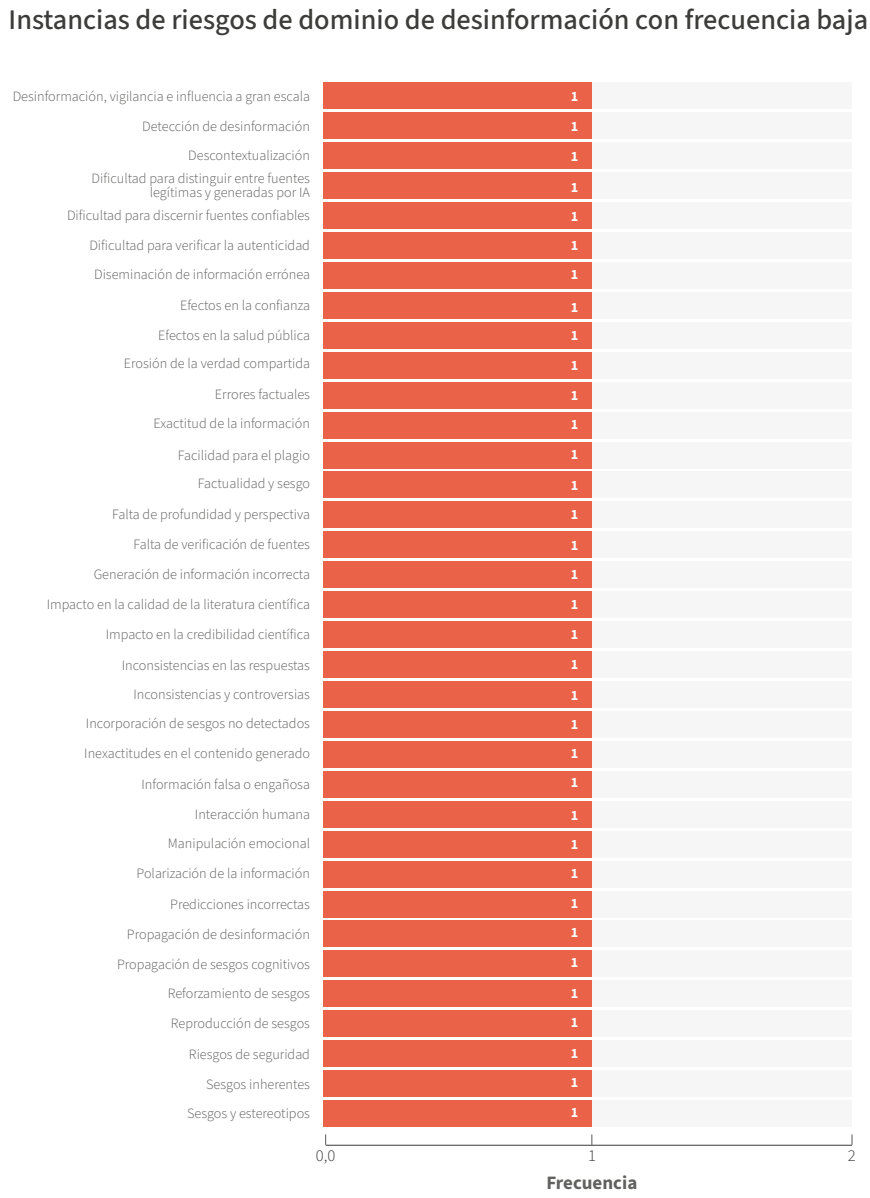
Instancias de riesgos del dominio de desinformación con frecuencia alta



Fuente: Elaboración propia

En forma complementaria, se muestran los riesgos del dominio de desinformación con frecuencia baja en la Figura 7.

Figura 7
Riesgos del dominio de desinformación con frecuencia baja



Fuente: Elaboración propia.

4.2 Dominios de Riesgos del Uso de la Inteligencia Artificial en la Investigación Científica

Los principales dominios de riesgo en el uso de IA en investigación científica se visualizan en la figura 8, en la cual se listan los principales dominios de riesgos que potencialmente se pueden presentar al integrar tecnologías de inteligencia artificial en los procesos de investigación científica.

Figura 8

Dominios de riesgos del uso de la inteligencia artificial en la investigación científica

Fuente: Elaboración propia

4.2.1 Integridad de Datos y Resultados

La integridad de datos se considera el core de la fundamentación epistemológica de la investigación científica. En el contexto actual, la integración de herramientas impulsadas por IA en los procesos

asociados a la investigación ha dado lugar a nuevos riesgos que amenazan este principio fundamental por la naturaleza sofisticada que en ocasiones se hace difícil de detectar. Por otro lado, este concepto abarca las amenazas y vulnerabilidades que comprometen la exactitud, integridad, validez y fiabilidad de los datos científicos, así como los resultados generados por sistemas de inteligencia artificial en la investigación científica. Este riesgo afecta todo el ciclo de vida de los datos, incide en los procesos analíticos, influye en la validez de las conclusiones y compromete el rigor científico. Entre los componentes críticos a considerar se encuentran la calidad de los datos de entrada, los procesos de transformación y análisis, la validez de los resultados esperados, la reproducibilidad de los hallazgos y la integridad de los sistemas de verificación y validación.

Entre las manifestaciones específicas de este riesgo se incluyen los sesgos en los conjuntos de datos de entrenamiento, las correlaciones engañosas, la manipulación inadvertida de información, las distorsiones de contenido no detectadas y los problemas de reproducibilidad. Complementariamente, entre los impactos de este dominio de riesgo se pueden mencionar el compromiso de la validez científica, la generación de conclusiones erróneas, la afectación de decisiones basadas en evidencia, la disminución de la confiabilidad de la investigación y la distorsión del conocimiento acumulado. Otras fuentes de información, enuncian como instancias del dominio de riesgos de integridad de datos a la fabricación de datos con IA, falsificación sofisticada de resultados experimentales y manipulación de imágenes en disciplinas de responsabilidad social alta como el área de la salud.

En cuanto a la fabricación de datos con IA, se han identificado diferentes técnicas, entre las cuales se pueden nombrar la generación de datasets experimentales de manera sintética, la interpolación maliciosa, la simulación sesgada, la síntesis de datos longitudinales. En lo referente a la instancia de falsificación sofisticada de resultados experimentales se han identificado técnicas como ajuste algorítmico de distribuciones, eliminación selectiva de datos que contradicen hipótesis a través de la IA, generación de datos de control de forma sintética. Finalmente, en el

contexto de la instancia del dominio de manipulación de imágenes científicas mediante IA se identifican riesgos en disciplinas donde la identificación visual es fundamental para la validación de hipótesis; entre las técnicas identificadas se enuncian la generación de imágenes microscópicas sintéticas, alteración de resultados a través de la IA, modificación de imágenes de resonancia magnética con el fin de simular patologías, creación de imágenes histológicas falsas no identificables de tejidos reales.

4.2.2 Transparencia y Explicabilidad

Este concepto engloba un conjunto de amenazas y vulnerabilidades asociadas a la falta de claridad en la parte algorítmica de los sistemas de IA, a su interpretabilidad y la ausencia de seguimiento y control en los sistemas de inteligencia artificial empleados en los procesos de investigación científica. Estos factores afectan la comprensión, verificación y validación de los procesos y resultados. En cuanto a la naturaleza del riesgo, destacan la falta de claridad en los procesos internos, la complejidad algorítmica desconocida, la ausencia de acceso a la trazabilidad de las decisiones automatizadas, la dificultad en la ejecución de auditorías y los problemas en la verificación integral. Entre los componentes críticos se incluyen la falta de acceso al conocimiento interno del algoritmo, la dificultad en la interpretación de los datos, la imposibilidad de realizar pruebas a los algoritmos que sustentan los sistemas, la falta de comprensión de las decisiones algorítmicas y la carencia de manuales técnicos y documentación del sistema.

Las manifestaciones de este dominio incluyen el uso de modelos de caja negra (solo se prueban entradas y salidas sin verificación de los algoritmos) en el diseño de los sistemas, la ausencia de condiciones que permitan comprender su funcionamiento interno, la falta de procesos para justificar los hallazgos, la complejidad en la implementación que dificulta los procesos de verificación y la predominancia de sistemas sin documentación, lo que impide la aplicación de pruebas de caja blanca en los sistemas de inteligencia artificial. En cuanto a los impactos, estos incluyen el compromiso de la rendición de cuentas respecto al conocimiento generado ante la comunidad científica, la falta de condiciones para

los procesos de verificación, la limitación en la reproducibilidad del conocimiento, la pérdida de confianza en los resultados y la generación de obstáculos para la colaboración interdisciplinaria. En esta categoría de riesgos, se puede ubicar lo concerniente al sesgo algoritmo, materializado en la falta de explicabilidad de los algoritmos que implementan los modelos de IA.

Un concepto relacionado con la transparencia y la explicabilidad es la opacidad asociada al diseño interior de los algoritmos que implementan los sistemas de IA, lo cual genera riesgos para la transparencia científica en cuanto a que afecta la validación, la potencialidad de réplica de los métodos y el avance del conocimiento científico. En forma específica, se han identificado tres categorías de opacidad, la arquitectónica, la de datos y la algorítmica. La opacidad arquitectónica se puede definir como la dificultad al comprender el procesamiento interno de las redes neuronales al procesar millones de parámetros. Por otro lado, la opacidad de datos especificada como la falta de transparencia en los conjuntos de datos utilizados para entrenar los modelos. Finalmente, la opacidad algorítmica explicada como la falta de fundamentación (documentación de las decisiones) en los ajustes y procesos de optimización, lo cual dificulta la reproducibilidad de los sistemas en otros contextos.

En lo referente a la transparencia, los procesos de investigación en explicabilidad de IA han identificado tres niveles de transparencia para garantizar la integridad científica, el primero es la transparencia algorítmica definida como la comprensión del diseño interior del modelo, incluyendo la arquitectura, los algoritmos de implementación y el proceso de entrenamiento con el origen de los datos. El segundo nivel corresponde a la transparencia de interacción, la cual se interpreta como la claridad en cuanto a la forma como el sistema interactúa con los datos de entrada y como produce los resultados particulares. El tercer nivel hace referencia a la transparencia social explicada como la comprensión del contexto social, ético y de operación en el cual el sistema de IA se ejecutará.

Una fundamentación especial en esta sección se debe hacer al sesgo de datos de entrenamiento, el cual se describe como los sesgos que se incorporan al conjunto de datos de entrenamiento de los sistemas de IA, los cuales pueden tener un impacto potencial en la validez científica. Particularmente, como instancias de esta tipología de sesgos se pueden considerar los sesgos de selección, los sesgos históricos, los sesgos de confirmación y los sesgos de disponibilidad.

4.2.3 Responsabilidad y Autoría

El dominio de riesgo de responsabilidad y autoría se define como el conjunto de amenazas y vulnerabilidades relacionadas con la atribución, asignación y delimitación de la responsabilidad intelectual, ética y legal en los procesos de investigación científica impulsados, asistidos o apoyados por inteligencia artificial. Este tipo de riesgo se manifiesta a través de la ambigüedad en la atribución de las contribuciones, la falta de asunción de responsabilidad ante errores o impactos derivados del uso indebido del conocimiento generado por sistemas de inteligencia artificial, la dificultad para identificar de manera precisa la propiedad intelectual y los problemas en la determinación de la autoría.

Entre los componentes críticos de este dominio de riesgo se puede enumerar la determinación de la contribución creativa o la determinación de que segmentos del conocimiento tienen origen en los humanos y cuales son generados por los sistemas de inteligencia artificial, asignar la responsabilidad por los impactos de los resultados generados por las inteligencias artificiales, identificación clara del crédito académico, claridad en la atribución de la propiedad intelectual, delimitación de responsabilidad por los errores que puedan tener impacto en la responsabilidad social. En forma complementaria, las manifestaciones de este riesgo se pueden dar en escenarios de disputas sobre autorías y contribuciones, litigios sobre atribución de descubrimientos, asignación de responsabilidad por fallos, conflictos por la propiedad de los resultados, ambigüedad por la asignación de los créditos académicos e investigativos en la autoría.

Por otro lado, una unidad de análisis importante en este tópico son los niveles de contribución de los sistemas de IA identificados en los procesos de investigación científica; particularmente se especifican a nivel de herramienta, nivel de colaborador, nivel de co-autor y nivel de investigador principal. En el nivel de utilización de la IA como herramienta la responsabilidad completa se atribuye al investigador. En el escenario de nivel de colaborador la responsabilidad es compartida entre el investigador y la casa comercial del modelo de IA de acuerdo a las normas regulatorias y a las cláusulas de uso del modelo. En el nivel de coautor se materializa un dilema ético acerca de la atribución de la autoría. El nivel de investigador principal de la IA se presenta un conflicto complejo respecto a la intervención humana en los procesos de verificación, validación y control de los resultados generados por la IA.

4.2.4 Impacto Social y Ético

La integración de la IAG en la investigación científica genera impactos sociales más allá del contexto académico; particularmente, ampliando brechas relacionadas con la equidad, la justicia y el acceso al conocimiento. En forma específica, se pueden presentar inequidades en el acceso, en el escenario en donde instituciones con recursos limitados tienen menos acceso a herramientas avanzadas de IA y equipos con capacidad de procesamiento suficiente para el entrenamiento de sistemas de IA. Por otro lado, los conjuntos de datos usados para entrenar los modelos de IA pueden no llegar a representar poblaciones marginales, presentando resultados posteriores que no incluyen este tipo de comunidades. Finalmente, en lo referente a la inequidad de los beneficios, los resultados de investigaciones con IA tienden a beneficiar a grupos y poblaciones ya privilegiadas.

A manera de síntesis, el dominio de riesgo Impacto social y ético se define como el conjunto de amenazas y vulnerabilidades asociadas con la perpetuación y ampliación de desigualdades existentes en la investigación y que potencialmente pueden ampliarse a través de la utilización de los sistemas de inteligencia artificial para apoyar la investigación científica. Adicionalmente, también se puede

manifestar por medio del mal uso que se le den a los hallazgos de la investigación generados con la ayuda de la inteligencia artificial; también, se puede manifestar en la ausencia de estrategias de verificación humanas que complementen procesos de toma de decisiones críticas automatizados. Finalmente, se pueden empoderar monopolios y concentración de poder en organizaciones con infraestructura de acceso y procesamiento de IA avanzada.

4.2.5 Calidad Científica

El dominio de riesgo de calidad científica se refiere a la disminución de los estándares de rigor, precisión y validez del conocimiento generado, comunicado y utilizado en el ámbito académico y científico. Este riesgo se manifiesta cuando la inteligencia artificial generativa produce contenido sin una fundamentación basada en hechos verificables, lo que puede dar lugar a alucinaciones que se difunden sin pasar por los procesos de verificación y validación propios del quehacer científico.

Uno de los principales impactos en la calidad científica es la generación de referencias bibliográficas inexistentes, es decir, citas que, al ser verificadas, no se encuentran en bases de datos ni en repositorios académicos. Además, este dominio de riesgo abarca otros problemas, como la dependencia excesiva de la automatización sin procesos adecuados de verificación y validación humana, la posible pérdida de habilidades investigativas fundamentales debido a la sobre dependencia de herramientas basadas en inteligencia artificial y el aumento de la producción científica masiva con bajos estándares de calidad.

En forma complementaria, la facilidad de uso de herramientas impulsadas por IA puede contribuir a prácticas que comprometan el rigor científico, como la aplicación de algoritmos de IA sin conocer y comprender el diseño interior de dichos algoritmos (sesgo algorítmico), lo cual puede a conducir a automatizaciones sin conocimiento de las limitaciones y supuestos de los algoritmos en su diseño. Adicionalmente, se pueden llegar a escenarios en los cuales se aceptan los resultados de la IA sin procesos de verificación y validación por parte de usuarios expertos disciplinares.

4.2.6 Desinformación Científica

Se define como el conjunto de amenazas y vulnerabilidades asociadas a la creación y difusión intencional de contenido científico falso o generado sin estrategias de verificación y validación mediante sistemas de inteligencia artificial, con el propósito de distorsionar el conocimiento, manipular la percepción o afectar la integridad del conocimiento científico. Entre sus características fundamentales, se puede señalar que la naturaleza del riesgo implica una intencionalidad deliberada, el uso de técnicas sofisticadas de difusión, un impacto sistémico en la generación de conocimiento y una alta dificultad en la detección de alucinaciones debido a la aparente fiabilidad del contenido.

Las manifestaciones de este tipo de riesgo pueden incluir la publicación de artículos académicos sin verificación ni validación adecuadas, la generación de datos y resultados sintéticos, la manipulación de imágenes científicas, la inclusión de referencias y citas inexistentes, así como la creación de identidades académicas falsas.

4.3 Dominio de Riesgos de Desinformación

4.3.1 Definición

El dominio de riesgos de desinformación en investigación científica se puede definir como el conjunto de amenazas y vulnerabilidades relacionadas con la generación, amplificación y difusión de información científica falsa o engañosa mediante el uso de sistemas de inteligencia artificial, que comprometen la integridad, credibilidad y calidad del proceso de investigación científica. Estos riesgos pueden impactar de manera negativa a través de la erosión de la confianza científica, decisiones basadas en información incorrecta, fragmentación del conocimiento científico debido a la generación de alucinaciones y la ausencia de estrategias de verificación y validación del conocimiento que se produce por las inteligencias artificiales generativas.

4.3.2 Características de los Riesgos del Dominio de Desinformación

Las características principales del dominio de riesgo de desinformación en investigación y comunicación científica se visualizan en la Figura 9, en la cual se especifican características como automatización y escalabilidad, sofisticación técnica, persistencia y mutabilidad, impacto sistémico, complejidad de detección, efecto multiplicador, resistencia a controles, impacto longitudinal.

Figura 9

Características de los riesgos de desinformación

Fuente: Elaboración propia.

4.3.2.1 Automatización y escalabilidad

Esta característica hace referencia a la generación masiva de contenido engañoso, falso o inventado, así como a la producción de conocimiento a gran escala mediante campañas publicitarias.

Complementariamente, la velocidad de propagación de este contenido es exponencial en comparación con los métodos tradicionales de difusión. Además, este riesgo puede manifestarse de forma recurrente en diversos contextos de comunicación e investigación científica. Finalmente, también pueden reproducirse patrones que no han sido suficientemente identificados.

En el contexto de la desinformación, se puede definir la automatización como la capacidad de sistemas de IA para ejecutar todo el proceso de creación, distribución y gestión de contenido falso sin intervención humana continua. Lo anterior incluye, generación de contenido, adaptación contextual, distribución coordinada, gestión de interacciones, evasión adaptativa.

4.3.2.2 Sofisticación Técnica

La sofisticación técnica se refiere a la capacidad avanzada de los sistemas de inteligencia artificial para generar, manipular, distribuir y reproducir información científica engañosa o falsa sin procesos adecuados de verificación y validación del conocimiento, alcanzando altos niveles de complejidad, precisión y aparente fiabilidad. Entre los aspectos esenciales de esta característica se incluyen las capacidades de generación, las técnicas de manipulación, las características tecnológicas y los mecanismos de engaño.

Otra corriente de información, respecto a la sofisticación técnica, especifica que la IAG produce conocimiento de alta fidelidad, como por ejemplo los deepfakes, los cuales en ocasiones son difíciles de diferenciar por los humanos y los sistemas de detección con respecto al contenido real. Esta sofisticación técnica reduce la posibilidad de identificar contenido engañoso y al mismo tiempo puede aumentar la credibilidad percibida por un usuario tradicional no experto disciplinar.

4.3.2.3 Persistencia y Mutabilidad

Representa la capacidad de los sistemas de inteligencia artificial para mantener, adaptar y transformar contenido falso o engañoso de manera continua, evadiendo los mecanismos de detección, verificación, validación y control, sin perder la intencionalidad desinformativa. Entre los aspectos esenciales de esta característica se incluyen las dimensiones de persistencia, los mecanismos de mutación, la caracterización tecnológica y el impacto informativo.

Por otro lado, el contenido generado por la IA puede ser modificado fácilmente (mutabilidad) para adaptarlo a diferentes audiencias, lo cual dificulta su trazabilidad y neutralización. Una vez que el contenido se difunde por los medios digitales, su eliminación total es casi imposible (persistencia), dejando una trazabilidad digital duradera.

4.3.2.4 Impacto Sistémico

La complejidad de detección comprende los desafíos técnicos, metodológicos y cognitivos asociados a la identificación, verificación y validación de contenido falso, inventado o engañoso generado por inteligencias artificiales generativas en el ecosistema de comunicación e investigación científica. Esta característica abarca aspectos esenciales como las dimensiones técnicas, las barreras metodológicas, los desafíos cognitivos y la caracterización del contenido falso.

En forma complementaria, la desinformación impulsada por IA, tiene la posibilidad de influir en procesos de opinión pública como las elecciones, mercados financieros, salud pública, afectando los procesos de toma de decisiones a nivel colectivo y ampliando brechas sociales y políticas en diferentes escenarios.

4.3.2.5 Complejidad de Detección

La complejidad de detección comprende los desafíos técnicos, metodológicos y cognitivos asociados a la identificación, verificación y validación de contenido falso, inventado o engañoso generado por inteligencias artificiales generativas en el ecosistema

de comunicación e investigación científica. Esta característica abarca aspectos esenciales como las dimensiones técnicas, las barreras metodológicas, los desafíos cognitivos y la caracterización del contenido falso.

Por otro lado, la competencia entre los sistemas de IA y los sistemas de detección de IA se ha convertido en un proceso de desarrollos técnicos continuos; por un lado los modelos de IA están empleando algoritmos más sofisticados para evitar la detección de IA, lo cual conduce a que los sistemas de detección de IA empleen arquitecturas más complejas, efectivas y proactivas generando de esta manera ciclos de detección difíciles de entender para incrementar la complejidad en los procesos de evitar la detección.

4.3.2.6 Efecto Multiplicador

Representa la capacidad de los sistemas de inteligencia artificial para amplificar, propagar, replicar y reproducir contenido falso o engañoso de manera exponencial en el ecosistema de comunicación e investigación científica. Esta característica abarca aspectos esenciales como los mecanismos de propagación, las dimensiones de la amplificación, las estrategias de replicación, la caracterización tecnológica y el impacto de la información difundida.

Adicionalmente, los sistemas de IA no solo generan conocimiento, sino que también optimizan los procesos de distribución a través de algoritmos de recomendación en las redes sociales, lo cual amplifica la velocidad, focaliza la población destinataria y el alcance de la desinformación potencialmente se vuelve más efectiva, dirigiéndola a usuarios específicos con una precisión psicológica mayor con respecto a los medios tradicionales.

4.3.2.7 Resistencia a Controles

Esta característica representa la capacidad de los sistemas de inteligencia artificial para evadir, adaptarse y sortear los controles de verificación, validación y mitigación de alucinaciones en el ecosistema científico. Para su análisis, es fundamental considerar aspectos como los mecanismos de evasión, las estrategias de superación, las dimensiones de adaptabilidad y las características técnicas.

Los controles tradicionales, como la moderación manual o las políticas de plataformas, no son suficientes ante el volumen y la velocidad del conocimiento generado por la IA. La naturaleza descentralizada de la IAG dificulta la aplicación de estrategias regulativas en el contexto de la desinformación.

4.3.2.8 Impacto Longitudinal

La capacidad de la desinformación producida por la inteligencia artificial para generar efectos acumulativos, persistentes a través del tiempo en el contexto científico. Para analizar esta característica se deben tener en cuenta aspectos relacionados con dimensiones temporales, mecanismos de reproducción, efectos en la estructura del ecosistema científico, impacto en la producción científica.

La exposición constante a información falsa o engañosa, puede llegar a alterar la memoria y las creencias a largo plazo. Este impacto longitudinal puede potencialmente afectar la formación de la opinión pública y la capacidad de una sociedad para operar con base en una memoria de conocimiento compartida susceptible de ser permeada por procesos de desinformación.

4.4 Subdominios del Dominio de Riesgo de Desinformación

Los subdominios del dominio de riesgos de desinformación en la comunicación y la investigación científica se visualizan en la figura 10. En forma particular, en la figura 10 se especifican los subdominios del riesgo de desinformación en la ciencia como son la generación de contenido falso, amplificación de la desinformación, la suplantación y falsificación, manipulación de revisión por pares, distorsión metodológica.

Figura 10
Subdominios del riesgo de desinformación

Fuente: Elaboración propia.

4.4.1 Generación de Contenido Falso

Este subdominio representa el proceso sistemático de generación de conocimiento no verificado por parte de los sistemas de inteligencia artificial, los cuales pueden producir textos, resúmenes, artículos científicos, datos, imágenes y referencias con altos niveles de fiabilidad, haciéndolos parecer originales y atribuidos a autores humanos. Para su análisis, es fundamental abordar temas como la tipología del contenido falso, los mecanismos de generación, las características técnicas y la precisión en la contextualización. Este fenómeno representa un desafío para la integridad de la información científica, ya que los sistemas de inteligencia artificial generan contenidos cada vez más convincentes y difíciles de identificar.

Por otro lado, la generación de contenido falso a través de la IA, abarca la creación automatizada de materiales que simulan procesos de investigación fiables, pero carecen de base empírica o metodología válida. Entre las categorías de contenido falso generado en el contexto de la investigación científica se pueden

enumerar la generación de artículos de investigación sintéticos completos, preprints y working papers sintéticos, materiales suplementarios falsos.

4.4.2 Amplificación de Desinformación

Este subdominio no solo abarca la generación de conocimiento inexacto, engañoso o falso, sino también su refuerzo, ampliación, diseminación y replicación, lo que incrementa su impacto tanto en la comunidad científica como en la sociedad en general. Entre los aspectos esenciales de este subdominio se incluyen la difusión a gran escala, el refuerzo algorítmico, el impacto en la toma de decisiones y la retroalimentación engañosa en los modelos de inteligencia artificial. Además, resulta fundamental la implementación de estrategias de verificación y validación del conocimiento por parte de los humanos, con el objetivo de evitar la amplificación de contenido inexacto o alucinaciones generadas por la IA.

Adicionalmente, la amplificación representa el subdominio que transforma contenido falso en procesos virales con impacto sistémico y de alta reproducibilidad. Este proceso opera mediante la explotación coordinada de algoritmos, redes sociales y los sesgos cognitivos humanos de acuerdo a su formación y creencias. Entre los mecanismos de amplificación algorítmica se encuentran las redes de bots científicos, la explotación de algoritmos de recomendación, cascadas de información y efecto multiplicador, amplificación Cross-platform, targeting de comunidades científicas.

4.4.3 Suplantación y Falsificación

Este subdominio se refiere a la generación intencional de comunicaciones o contenidos científicos que imitan de manera fraudulenta la identidad, las credenciales, el estilo y la metodología de autores, utilizando sistemas de inteligencia artificial con el propósito de engañar, manipular, extorsionar o desinformar. Para su análisis, es necesario estudiar aspectos como las modalidades de suplantación, las funcionalidades de las herramientas empleadas para la falsificación, los impactos potenciales de su materialización, así como las estrategias de detección y los procesos de mitigación

de la suplantación y falsificación. En forma complementaria, se instancian riesgos en este subdominio relacionados con la imitación de estilos de escritura académica, creación de perfiles falsos de investigadores, falsificación de credenciales y manipulación de afiliaciones institucionales.

De otra manera, este subdominio abarca la creación y utilización de identidades científicas falsas o suplantadas para otorgar credibilidad a la desinformación, explotando el principio de autoridad que es común en la comunidad científica. Entre las instancias de este subdominio se pueden enumerar la creación sintética de perfiles de investigadores, la suplantación de investigadores legítimos, creación de instituciones y journals falsos.

4.4.4 Distorsión Metodológica

La distorsión metodológica a través de sistemas de inteligencia artificial se refiere a la manipulación, intencional o accidental, de los procesos, protocolos y estándares de la investigación científica, lo que afecta la integridad, validez y reproducibilidad de los resultados, así como su difusión. Para el análisis de este subdominio, es fundamental considerar las distintas modalidades de distorsión metodológica, entre las que se incluyen la manipulación de datos, la alteración de análisis estadísticos, la simulación de experimentos y la fabricación de evidencia científica.

En el contexto de este subdominio, se deben considerar los sistemas de inteligencia artificial capaces de generar datos sintéticos, los algoritmos y las aplicaciones con potencial para amplificar sesgos, así como las aplicaciones impulsadas por IA para la visualización de datos, que pueden distorsionar los gráficos generados. La materialización de los riesgos asociados a este subdominio puede afectar la integridad científica, disminuir la confianza en las publicaciones dentro del proceso de revisión por pares y propiciar la producción de conocimiento científico potencialmente erróneo, inexacto o engañoso.

En este subdominio, la IA es utilizada para crear apariencia de rigor científico mientras se afectan fundamentos metodológicos primordiales para la validez del conocimiento científico. Entre

las instancias de este subdominio se encuentra la fabricación de metodología aparentemente rigurosa, P-Hacking y HARKing automatizados, manipulación de análisis de datos, barreras para la reproducibilidad del conocimiento, reproducibilidad conceptual alterada.

4.4.5 Manipulación de Revisión por Pares

Este subdominio se refiere a la alteración intencional o sistemática de los procesos tradicionales de evaluación científica en la revisión por pares, tanto en congresos como en revistas de investigación. Estas alteraciones se llevan a cabo mediante el uso de tecnologías de inteligencia artificial con el propósito de influir, sesgar o comprometer la integridad de los procesos de revisión científica. Entre los principales riesgos asociados a este subdominio se encuentran la generación sintética de revisiones, la suplantación de revisores, la manipulación de sistemas de gestión editorial y la alteración de los criterios de evaluación.

Los impactos potenciales de los riesgos asociados a este subdominio incluyen la disminución de la credibilidad del sistema de revisión por pares, la publicación de investigaciones con un rigor científico reducido, la amplificación de sesgos sistemáticos en la comunicación científica y la desviación de recursos hacia investigaciones con fundamentos insuficientes.

Adicionalmente, el proceso de revisión de pares se constituye en el mecanismo fundamental de control de calidad de la ciencia. La integración de la IA en este proceso, ha creado vulnerabilidades que comprometen la integridad del conocimiento científico publicado. Entre las instancias de este subdominio se pueden enumerar la generación de revisión por pares sintéticos, la explotación de sistemas de revisión por pares, erosión de la confianza de la revisión por pares.

4.5 Enfoque HITL (Human-in-the-Loop)

El enfoque HITL (Human-in-the-Loop) es un paradigma aplicado tanto al diseño de sistemas de inteligencia artificial con intervención humana como a la verificación y validación de los resultados generados por estas tecnologías en tiempo real (Fill et al., 2024). Este enfoque integra la supervisión humana a lo largo del ciclo de diseño, implementación, verificación, validación y difusión del conocimiento, con el objetivo de mejorar la precisión, fiabilidad, interpretabilidad, confiabilidad e integridad de los resultados de la inteligencia artificial en el contexto de los procesos de comunicación e investigación científica. Su aplicación es especialmente valiosa en dominios de alta complejidad y responsabilidad social, como la investigación científica, la medicina, la toma de decisiones en el ámbito judicial y la gestión semiautomatizada de recursos humanos.

4.5.1 Niveles de Interacción Humana en HITL

Los niveles básicos de interacción humana en HITL se visualizan en la Figura 11, en la cual se identifican los roles que pueden desempeñar los humanos en los procesos de verificación y validación del conocimiento generado por la IA. Particularmente, en la figura se visualizan las categorías de intervención humana en HITL como son la supervisión pasiva, la verificación activa, la retroalimentación iterativa, la cocreación interactiva, la intervención preventiva, la supervisión especializada, auditoria sistemática.

Figura 11
Niveles de interacción humana en HITL

Fuente: Elaboración propia.

4.5.1.1 Supervisión Pasiva

Este nivel de interacción humana implica la supervisión sin una intervención inmediata. El usuario actúa como un observador que verifica los resultados una vez generados. Entre sus características, se destaca que la revisión se realiza en la etapa de postprocesamiento, permitiendo la identificación de errores evidentes y una verificación general de coherencia. Por otro lado, la supervisión pasiva representa el nivel de intervención humana menos intrusivo en tiempo real. En este modelo, frecuentemente denominado Human-on-the-loop (HOTL), el sistema de IA opera con un alto grado de autonomía, ejecutando sus tareas y tomando decisiones sin requerir una supervisión humana en cada paso. El rol del humano es el de un supervisor o vigilante que monitorea el

rendimiento del sistema desde una posición externa al proceso de decisión principal.

La intervención es reactiva y se produce solo en circunstancias excepcionales: cuando el sistema presenta alguna falla que no puede resolver, cuando el nivel de confianza cae por debajo del nivel preconfigurado de confianza, cuando se activa una alarma de seguridad o cuando el supervisor humano detecta un comportamiento no normal. Entre las características clave se pueden enunciar una alta autonomía de la IA, intervención asincrónica y basada en excepciones, rol de monitorización, equilibrio entre autonomía y seguridad.

4.5.1.2 Verificación Activa

En este nivel de interacción humana, los usuarios realizan verificaciones específicas sobre los resultados generados por la inteligencia artificial antes de su publicación o uso. Su caracterización incluye la validación de fuentes, la comprobación de hechos y el contraste con referentes teóricos de la literatura científica más reconocida y con mayores índices de citación.

Adicionalmente, en este paradigma el sistema de IA no tiene autonomía total para completar su proceso. Su flujo de trabajo está diseñado con puntos de verificación deliberados en puntos críticos. En estos puntos, el sistema detiene su ejecución y solicita una acción humana como una aprobación, una corrección, una selección entre varias opciones o una confirmación final. La decisión humana no es una opción sino un requisito indispensable para continuar con el proceso. Entre las características clave se especifican los puntos de intervención deliberados, la decisión humana como condición, enfoque en la seguridad y la responsabilidad, control humano explícito.

4.5.1.3 Retroalimentación Iterativa

En este nivel de interacción humana, el mejoramiento continuo se hace evidente, ya que la verificación por parte de los usuarios permite mejorar gradualmente los resultados de la inteligencia artificial. Se caracteriza por la aplicación de correcciones

específicas, el refinamiento progresivo y el entrenamiento implícito del sistema. Adicionalmente, este nivel trasciende la simple verificación de un paso para transitar hacia un ciclo de aprendizaje dinámico y continuo. En el modelo de retroalimentación iterativa, la intervención humana no solo corrige y valida una decisión específica, sino que cada una de esas interacciones se registra y se utiliza para reentrenar, retroalimentar y mejorar el sistema de IA. El alcance central es que el sistema aprenda de su experiencia e interacciones en el mundo real, se adapte a nuevos patrones y reduzca la intervención humana paulatinamente como consecuencia de su aumento progresivo en la precisión y en la autonomía.

Entre las características clave de este nivel se pueden enumerar el enfoque en la adaptabilidad y mejora continua, uso sistemático de la retroalimentación, eficiencia a través de la inteligencia, alineación con preferencias humanas de acuerdo al contexto,

4.5.1.4 Cocreación Interactiva

En este nivel, los humanos y la inteligencia artificial trabajan en paralelo, con los humanos coordinando activamente el proceso de creación o análisis. La interacción se caracteriza por un diálogo constante, una toma de decisiones compartida y una segmentación de tareas basada en las fortalezas de cada componente del sistema. En forma complementaria, este nivel representa un cambio de paradigma fundamental, el humano y la IA se convierten en socios colaborativos en un proceso creativo o de resolución de problemas complejos. La interacción es dinámica, conversacional y con sinergia.

En cuanto al proceso, la IA no se limita a ejecutar tareas o solicitar aprobaciones; actúa como un generador de ideas para el pensamiento humano, un explorador de alternativas y un asistente que aumenta las capacidades cognitivas del humano. El objetivo no es la simple precisión o la eficiencia, sino la innovación, la exploración y el descubrimiento. Entre las características clave se pueden enumerar la relación de asociación, el proceso dialógico e iterativo, enfoque en la generación y exploración, aumento de la creatividad humana.

4.5.1.5 Intervención Preventiva

En este nivel, los humanos definen parámetros, restricciones y directrices antes de la generación de resultados por parte de la inteligencia artificial. Destaca la especificación de límites éticos, el establecimiento de criterios de verificación y la programación de alertas. Adicionalmente, este nivel representa un cambio de enfoque desde la corrección reactiva hacia el control proactivo. Particularmente, el rol principal del humano es el de actuar como arquitecto que diseña y construye un escenario seguro dentro del cual la IA puede operar. También, podrá anticipar posibles modos de fallos, riesgos y resultados indeseables, complementados con la implementación de marcos éticos antes y durante la operación del sistema para maximizar la prevención.

En forma complementaria, el arquitecto de un sistema de IA preventivo define los limites operativos, los umbrales de riesgo y los protocolos de emergencia para garantizar un funcionamiento seguro y fiable. Entre las características clave de este nivel se pueden especificar el enfoque proactivo en la mitigación de riesgos, la definición de limites operativos, el cumplimiento normativo y ético por diseño, requiere un conocimiento profundo del dominio y del sistema.

4.5.1.6 Supervisión Especializada

Este nivel de interacción involucra a expertos en dominios específicos para la verificación y validación de resultados en áreas de alta complejidad técnica. Se caracteriza por la intervención de especialistas, la aplicación constante de conocimiento disciplinar y una evaluación basada en la experticia. En forma complementaria, este nivel de interacción reconoce que no toda la intervención humana es igual.

La supervisión especializada de la simple verificación generalizada por parte de un operario; implica la participación de expertos de dominio cuyo conocimiento profundo, contextual y tácito es crucial para guiar, corregir y validar a la IA en tareas complejas y matizadas.

En estos escenarios, la IA es una herramienta excelente de análisis y procesamiento, pero al final la decisión depende del juicio humano experto, el cual el sistema no está en capacidad de replicar. Entre las características clave de este nivel se pueden enunciar la aportación de conocimiento tácito y contextual, la resolución de la ambigüedad, la validación de alto nivel, el enfoque en la calidad y fiabilidad del resultado final.

4.5.1.7 Auditoría sistemática

En este nivel de interacción se realiza una verificación y evaluación integral y metodológica de todo el ciclo de diseño, implementación y uso. Además, se examinan los resultados generados por la inteligencia artificial. Este tipo de auditoría se distingue por una documentación exhaustiva, trazabilidad completa y una evaluación rigurosa de sesgos y errores sistemáticos. Este nivel también es conocido como Humand-behind-the-loop (HBTL), es un nivel de interacción que se produce de forma retrospectiva, es decir, después de que el sistema de IA ha operado y tomado decisiones.

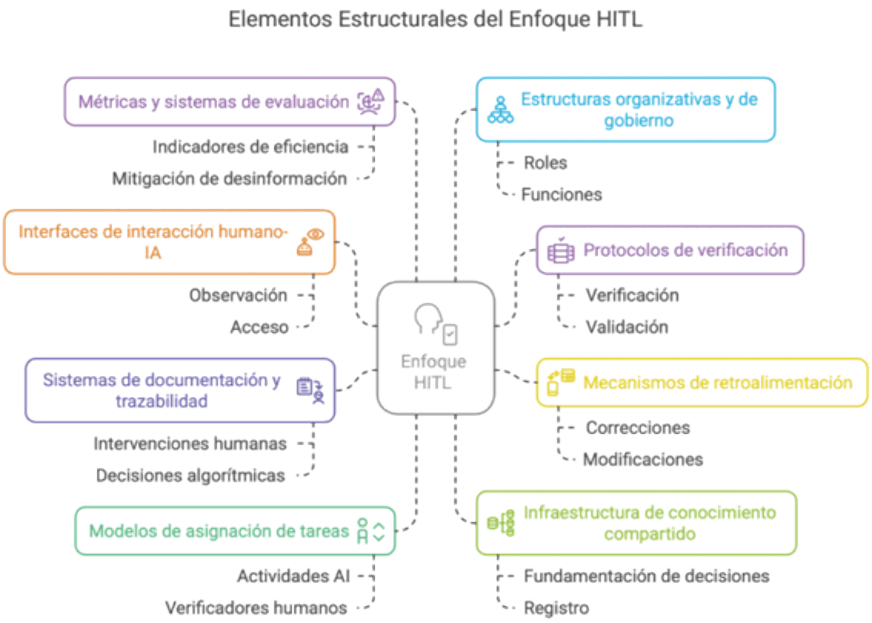
Adicionalmente, en este modelo el rol humano no es intervenir en tiempo real, sino analizar el rendimiento histórico del sistema para evaluar su eficiencia, detectar sesgos sistémicos, garantizar el cumplimiento normativo e identificar lecciones para futuras versiones mejoradas del modelo. Se convierte en una actividad de gobernanza, supervisión y rendición de cuentas fundamentalmente. Entre las características clave se pueden nombrar la actividad retrospectiva, el enfoque en la gobernanza y el cumplimiento, basado en datos y registros, impulsor de mejoras sistémicas.

4.5.2 Elementos Estructurales del HITL

El enfoque HITL necesita de la especificación de una estructura que le permita actuar como estrategia de verificación del conocimiento generado por las inteligencias artificiales generativas, los elementos estructurales del HITL se visualizan en la figura 12. Entre los elementos estructurales visualizados en la figura 12 se especifican las interfaces de interacción humano-IA, protocolos de verificación, mecanismos de retroalimentación, sistemas de documentación y trazabilidad, modelos de asignación de tareas,

infraestructura de conocimiento compartido, métricas y sistemas de evaluación, estructuras organizativas y de gobierno.

Figura 12
Elementos estructurales del enfoque HITL



Fuente: Elaboración propia.

4.5.2.1 Interfaces de Interacción Humano-IA

Se refiere a los puntos de contacto o interfaces mediante los cuales los humanos pueden observar, acceder, evaluar y modificar los procesos y resultados generados por la inteligencia artificial. De otra forma, este elemento es el punto de contacto directo entre el operador humano y el sistema de IA. Su diseño es crucial para garantizar una intervención humana eficiente y bien informada.

Una interfaz efectiva debe estar diseñada con el criterio de explicabilidad como fundamento, más allá de mostrar la decisión o respuesta final de la IA. Debe presentar el contexto y las evidencias que la IA utilizó para llegar a la conclusión final, incluyendo la visualización de los datos de entrada, las características clave que

influyeron en la predicción y el nivel de confianza del modelo. Adicionalmente, la interfaz debe incorporar una interfaz de anulación que permita al humano detener o modificar la decisión de la IA de forma clara y documentada.

4.5.2.2 Protocolos de Verificación

Hace referencia a procedimientos sistemáticos y estandarizados que guían a los humanos en la verificación y validación de los resultados generados por la inteligencia artificial. Adicionalmente, los protocolos de verificación definen las reglas y los procesos que determinan cuándo y cómo debe intervenir un humano, siendo esenciales para la eficiencia y la seguridad del sistema.

La intervención humana se activa mediante disparadores de revisión basados en el riesgo, que incluyen: un umbral de confianza bajo de la IA, la naturaleza de alto riesgo de la decisión, la detección de patrones novedosos o casos limite y el cumplimiento de requisitos regulatorios que exigen supervisión humana obligatoria. Para asegurar la calidad de la intervención, se implementan procesos de aseguramiento de la calidad para asegurar la consistencia del revisor humano.

4.5.2.3 Mecanismos de Retroalimentación

Sistemas o procedimientos que permiten integrar correcciones y modificaciones a los sistemas de inteligencia artificial tras la aplicación de estrategias de verificación y validación del conocimiento con el fin de mejorar la precisión de los resultados de la inteligencia artificial. De otra forma, los mecanismos de retroalimentación son el motor que impulsa la mejora del modelo de IA. Debe existir un mecanismo de captura de retroalimentación formal en la interfaz de revisión que registre la razón específica de la anulación o corrección humana. Esta retroalimentación capturada, que representa el juicio humano, se convierte en datos de entrenamiento de alta calidad que se reincorporan al modelo de IA para su reentrenamiento o ajuste, cerrando el bucle de mejora.

En los modelos en los cuales la IA toma la decisión automáticamente, se utiliza un muestreo post-decisión de las decisiones para una

revisión humana posterior, cuyo objetivo es detectar problemas sistemáticos o sesgos en el modelo y generar retroalimentación para la mejora continua.

4.5.2.4 Sistemas de Documentación y Trazabilidad

Sistemas que registran y le dan persistencia a las intervenciones humanas y las decisiones algorítmicas en el proceso de diseño, implementación, verificación, validación y uso del conocimiento generado, con el propósito de documentar la fundamentación de las decisiones en el contexto de los sistemas de inteligencia artificial. Estos sistemas son vitales para la rendición de cuentas, la auditoria y el cumplimiento normativo.

El componente central es el registro de auditoría, un registro no modificable y completo que captura la decisión original de la IA, la identidad del revisor humano, la marca de tiempo de la revisión, la razón de la anulación y la decisión final. Esta trazabilidad es fundamental para soportar el derecho a la explicación para las decisiones automatizadas dirigidas al usuario, permitiendo reconstruir el proceso de toma de decisiones.

4.5.2.5 Modelos de Asignación de Tareas

En el contexto del proceso de verificación y validación en el enfoque HITL, la asignación de tareas se refiere a la designación de actividades que serán realizadas por el sistema de inteligencia artificial y aquellas que estarán a cargo de los verificadores humanos. Adicionalmente, la asignación de tareas define la arquitectura del flujo de trabajo HITL, determinando la relación de autoridad y secuencia entre la IA y el humano. Existen tres modelos que pueden llegar a estructurar esta decisión, el primero es la revisión pre-decisión, el segundo es el muestreo post-decisión y el tercero es el enrutamiento por umbral de confianza.

4.5.2.6 Infraestructura de Conocimiento Compartido

Se refiere a sistemas diseñados para registrar la fundamentación de las decisiones, tanto a nivel de la intervención humana como en los sistemas de inteligencia artificial. De otra forma, este elemento

incluye los sistemas y procesos que permiten a los humanos compartir el conocimiento adquirido a través de sus interacciones con la IA, estandarizando el juicio humano.

Una infraestructura robusta incluye un repositorio de casos limite y anulaciones, una base de conocimiento centralizada que almacena y categoriza las decisiones humanas que anularon a la IA, sirviendo de material de referencia y entrenamiento. Este repositorio es la base para un programa de capacitación con documentación formal para los revisores humanos, cuyo objetivo es estandarizar las directrices de revisión y asegurar la consistencia en el juicio a través de directrices de revisión estandarizadas claras y accesibles.

4.5.2.7 Métricas y Sistemas de Evaluación

Mecanismos para medir los indicadores de eficiencia en las acciones destinadas a mitigar el riesgo de desinformación, tanto en el componente humano como en el componente algorítmico del sistema. De otra manera, la evaluación del rendimiento de un sistema HITL debe medir tanto la eficiencia de la IA como la calidad y la velocidad de la intervención humana. Las métricas clave se concentran en la interacción y el impacto de la intervención humana, entre las cuales se pueden especificar la tasa de acuerdo Humano-IA, la tasa de anulación, el tiempo promedio de revisión, la puntuación de auditoría de cumplimiento.

4.5.2.8 Estructuras Organizativas y de Gobierno

Definen roles, funciones, normas y procesos de toma de decisiones en el contexto del enfoque HITL. Adicionalmente, la gobernanza de HITL define la estructura de responsabilidad y la supervisión del sistema, siendo un componente esencial para el cumplimiento y la ética. Lo anterior comienza con una política de gobernanza de la IA que formaliza el marco HITL en el contexto de la organización.

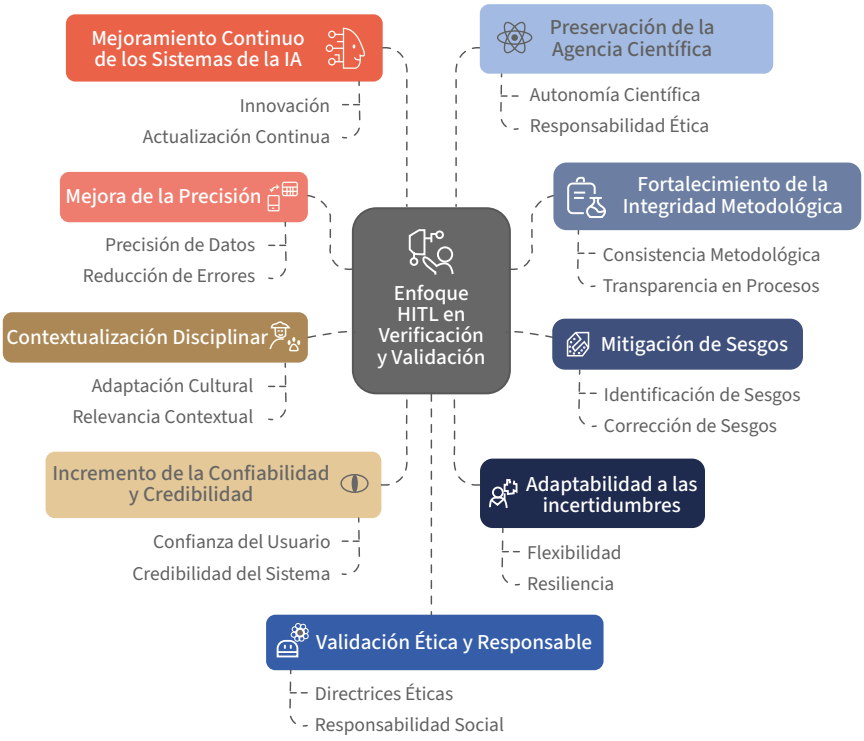
La estructura debe garantizar una asignación clara de roles y responsabilidades, definiendo la autoridad de decisión y la responsabilidad final. Los roles clave incluyen el propietario del sistema de IA, el experto de la materia, el revisor de calidad y el oficial de cumplimiento, cada uno de ellos con autoridad de decisión

específica. Es de alta importancia definir la responsabilidad por las decisiones tomadas y planear auditorias regulares internas y de terceros para asegurar la alineación a los protocolos y métricas.

4.5.3 Beneficios del Enfoque HITL para los Procesos de Verificación y Validación

Los beneficios clave del enfoque HITL en el proceso de verificación y validación, como estrategia para minimizar el impacto del riesgo de desinformación, se presentan en la Figura 13. Entre los beneficios visualizados y especificados en la figura 13 se cuentan la mejora de la precisión, fortalecimiento de la integridad metodológica, mitigación de sesgos, contextualización disciplinar, incremento de la confianza y la credibilidad, adaptabilidad a las incertidumbres, mejoramiento continuo de los sistemas de IA, preservación de la agencia científica, validación ética y responsable.

Figura 13
Beneficios del enfoque HITL en los procesos de verificación y validación



Fuente: Elaboración propia.

4.5.3.1 Mejora de la Precisión

La verificación humana puede identificar inexactitudes o conocimiento erróneo que los sistemas de IA pasan por alto, lo cual ayuda a mejorar la precisión y a controlar la creatividad en la generación de los resultados potencializando de esta manera la integridad del conocimiento producido por la inteligencia artificial generativa. En forma complementaria, los humanos contribuyen a la especificación del contexto, identificando cuando se presentan resultados fuera del contexto real requerido o cuando se hacen interpretaciones diferentes a las necesidades del usuario final.

4.5.3.2 Fortalecimiento de la Integridad Metodológica

Los verificadores humanos pueden corroborar que los métodos de investigación creados por la inteligencia artificial generativa sean pertinentes, robustos y alineados con los requerimientos de la disciplina. Además, pueden identificar las limitaciones de los métodos proporcionados por los sistemas de inteligencia artificial.

4.5.3.3 Mitigación de Sesgos

Los verificadores humanos pueden identificar sesgos en el conocimiento generado por la inteligencia artificial, especialmente cuando se pretende utilizar en procesos de comunicación e investigación científica. Además, pueden contribuir a la incorporación de perspectivas diversas y representativas de referentes científicos en el objeto de estudio. Finalmente, la verificación humana potencialmente puede contribuir a minimizar los sesgos generados a partir de los datos de entrenamiento de los sistemas de inteligencia artificial.

4.5.3.4 Contextualización Disciplinar

Los verificadores humanos y expertos en la disciplina pueden ajustar el conocimiento generado por los sistemas de inteligencia artificial según el contexto específico de la disciplina. Además, pueden complementar los resultados de la inteligencia artificial con avances, descubrimientos y desarrollos científicos producidos por

la comunidad académica después del período de entrenamiento del modelo de inteligencia artificial generativa.

4.5.3.5 Incremento de la Confiabilidad y Credibilidad

La verificación experta por parte de especialistas en la disciplina contribuye a aumentar la confianza y la fiabilidad de los resultados generados por los sistemas de inteligencia artificial. Además, los verificadores humanos pueden ayudar a mitigar los riesgos éticos específicos de cada disciplina científica.

4.5.3.6 Adaptabilidad a las Incertidumbres

Los verificadores expertos pueden identificar con mayor precisión los límites actuales del conocimiento científico, los cuales los sistemas de inteligencia artificial tienden a simplificar, generalmente basándose en los datos de entrenamiento. Además, pueden proporcionar recomendaciones o sugerencias para el uso del conocimiento generado por la inteligencia artificial, promoviendo un enfoque cauteloso y socialmente responsable según las particularidades de cada disciplina.

4.5.3.7 Mejoramiento Continuo de los Sistemas de IA

Las apreciaciones, hallazgos y correcciones identificados por los verificadores expertos humanos pueden ser retroalimentados e incorporados en los sistemas de inteligencia artificial para aumentar su precisión y garantizar su mejora continua. La detección de hallazgos mediante la verificación experta humana y su retroalimentación contribuye a optimizar la calidad y precisión de los futuros resultados generados por la inteligencia artificial.

4.5.3.8 Preservación de la Agencia Científica

Garantizan que las decisiones finales sobre la validez científica permanezcan bajo el control humano. Además, la verificación experta humana permite incorporar el conocimiento tácito y la experiencia, elementos que generalmente no están dentro del alcance de los datos de entrenamiento de la inteligencia artificial generativa.

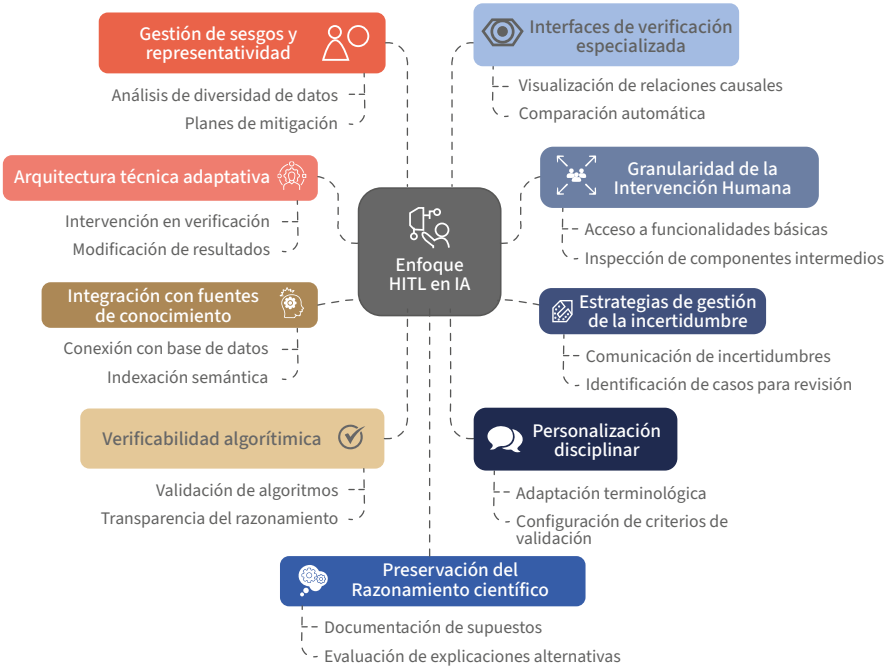
4.5.3.9 Validación Ética y Responsable

Los verificadores humanos pueden incorporar al sistema consideraciones éticas, de responsabilidad social, de beneficencia y de equidad en relación con el conocimiento generado por las inteligencias artificiales generativas. Además, la verificación experta humana permite controlar la propagación de información que aparenta ser fiable, pero carece de fundamentación científica, fenómeno conocido como alucinaciones.

4.5.4 Implicaciones del Enfoque HITL en los Sistemas de Inteligencia Artificial en el Contexto de la Comunicación e Investigación Científica

La implementación del enfoque HITL en sistemas de inteligencia artificial para apoyar los procesos de comunicación e investigación científica debe abordar los tópicos especificados en la figura 14. En forma específica, en la Figura 14 se visualizan los tópicos HITL a tener en cuenta en la comunicación e investigación científica como son la arquitectura técnica adaptativa, granularidad de la intervención humana, estrategias de la gestión de la incertidumbre, integración con fuentes de información de alta confianza, verificabilidad algorítmica, personalización disciplinar, gestión de sesgos y representatividad, interfases de verificación especializada, preservación del razonamiento científico, evaluación compartida humano – máquina.

Figura 14
Tópicos HITL en la comunicación e investigación científica



Fuente: Elaboración propia.

4.5.4.1 Arquitectura Técnica Adaptativa

El diseño de la arquitectura técnica del sistema de inteligencia artificial debe permitir la intervención humana en las distintas fases del desarrollo, con especial énfasis en la verificación y validación de los resultados generados por la inteligencia artificial generativa. Es fundamental considerar puntos de intervención humana en los diferentes niveles de la arquitectura, incluyendo interfaces y funcionalidades que permitan la modificación de los resultados, el ajuste de las variables de configuración y la capacidad del sistema para adaptar los flujos de procesamiento según las decisiones humanas.

4.5.4.2 Granularidad de la Intervención Humana

Nivel de detalle de los subsistemas, componentes y funcionalidades en los que los humanos pueden examinar, verificar, validar y modificar tanto los subsistemas como los resultados del sistema de inteligencia artificial. En este contexto, es fundamental equilibrar el acceso a funcionalidades básicas y abstracciones de alto nivel. Además, se debe garantizar el permiso para inspeccionar componentes intermedios de procesamiento y proporcionar control sobre las variables de configuración del modelo.

4.5.4.3 Estrategias de Gestión de la incertidumbre

Deben especificarse los procedimientos para comunicar y gestionar los niveles de confianza, ambigüedad e incertidumbre en los resultados y en los subsistemas intermedios de la arquitectura del sistema de inteligencia artificial. También es fundamental implementar mecanismos de identificación automática de los casos que requieran revisión humana. Además, el sistema de inteligencia artificial debe ser retroalimentado de forma continua con los avances científicos recientes y los desarrollos teóricos de los distintos campos disciplinares.

4.5.4.4 Integración con Fuentes de Conocimiento de Alta Confianza

Mecanismos de conexión del sistema de inteligencia artificial con bases de datos científicas, literatura académica validada, repositorios de preprints, repositorios de artefactos de investigación de universidades de alto reconocimiento a nivel mundial y otros recursos de alta confianza en el campo disciplinar. Se debe considerar la disponibilidad de acceso a repositorios científicos mediante APIs, la indexación semántica de artículos revisados por pares y la capacidad de generar citas precisas y verificables.

4.5.4.5 Verificabilidad Algorítmica

Se refiere a la capacidad del sistema para verificar, validar y examinar los algoritmos que implementan cada uno de los subsistemas de

la arquitectura de la inteligencia artificial, de manera similar a las pruebas de caja blanca. Además, es fundamental garantizar la persistencia de la fundamentación de las decisiones algorítmicas, la transparencia del proceso de razonamiento, la reutilización de componentes y el registro de subsistemas o componentes críticos del sistema de inteligencia artificial.

4.5.4.6 Personalización Disciplinar

Se refiere a la capacidad del sistema de IA para adaptarse a los paradigmas epistemológicos, metodológicos y comunicativos propios de cada disciplina. Además, se debe considerar la parametrización del sistema según el campo disciplinar, garantizar la adaptación a sus convenciones terminológicas específicas y planificar la configuración de los criterios de validación correspondientes.

4.5.4.7 Gestión de Sesgos y Representatividad

Se refiere a los procesos para identificar, documentar y desarrollar planes de mitigación de los sesgos potenciales originados en los datos de entrenamiento, los algoritmos o las interfaces del sistema de IA. De manera más detallada, esto implica realizar un análisis de la diversidad de los datos de entrenamiento, evaluar los impactos diferenciales, documentar las limitaciones del sistema y establecer mecanismos de verificación y corrección por parte de expertos humanos en cada disciplina.

4.5.4.8 Interfaces de Verificación Especializada

Se refiere a las herramientas, interfaces y funcionalidades diseñadas para facilitar a los expertos humanos en cada disciplina la verificación y validación de los contenidos generados por los sistemas de inteligencia artificial. Entre las consideraciones de este aspecto, se deben incluir la visualización de relaciones causales, la comparación automática con los avances científicos recientes y la validación de la existencia y precisión de citas y referencias.

4.5.4.9 Preservación del Razonamiento Científico

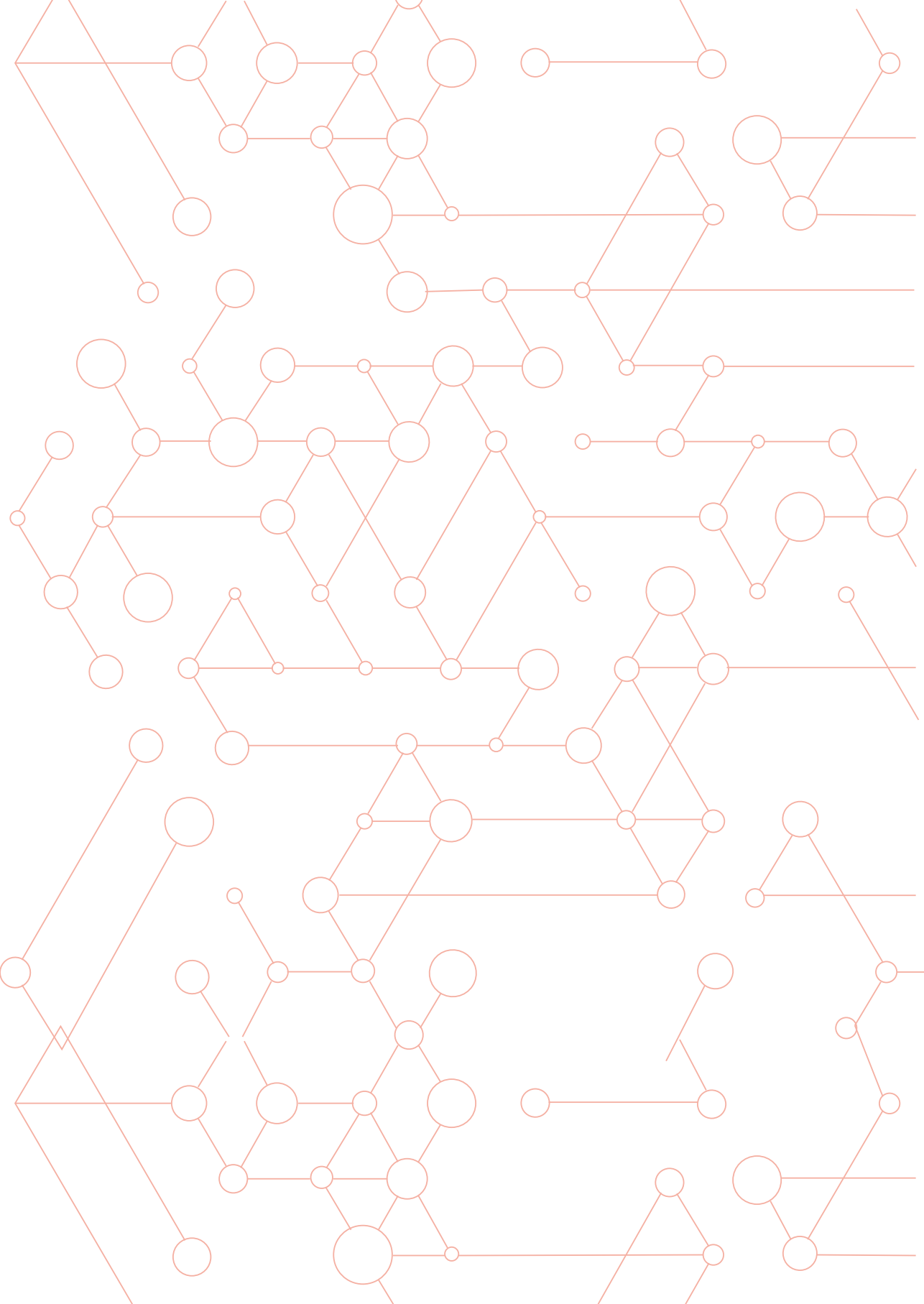
Se refiere a los mecanismos que garantizan la preservación de los procesos fundamentales del método científico en los sistemas híbridos de inteligencia artificial, especialmente en la integración de los subsistemas humano-IA. En este contexto, es fundamental considerar la especificación y justificación de los supuestos, la documentación y fundamentación de las limitaciones metodológicas, así como la definición de una estructura para la evaluación sistemática de explicaciones alternativas.

4.5.4.10 Evaluación Compartida Humano – Máquina

Hace referencia a los procesos para verificar, validar y evaluar sistemáticamente el desempeño del sistema autónomo de IA en comparación con la verificación humana o la evaluación integrada de los sistemas de IA y humanos. En este contexto, es fundamental gestionar métricas de precisión, tiempo de procesamiento, creatividad y otros indicadores específicos de cada disciplina científica.

CAPÍTULO V

CONSOLIDACIÓN DE LOS FUNDAMENTOS DE LOS RIESGOS DEL USO DE LA INTELIGENCIA ARTIFICIAL EN LA COMUNICACIÓN E INVESTIGACIÓN CIENTÍFICA. UNA VISIÓN ENFOCADA DESDE EL DOMINIO DE DESINFORMACIÓN



CAPÍTULO V

CONSOLIDACIÓN DE LOS FUNDAMENTOS DE LOS RIESGOS DEL USO DE LA INTELIGENCIA ARTIFICIAL EN LA COMUNICACIÓN E INVESTIGACIÓN CIENTÍFICA. UNA VISIÓN ENFOCADA DESDE EL DOMINIO DE DESINFORMACIÓN

Los resultados de este estudio destacan varios temas relacionados con los aspectos éticos del uso de la inteligencia artificial (IA) en la comunicación científica. Primero, se especifican los dominios en los cuales se pueden categorizar los riesgos asociados al uso de la IA; en particular, se identifican siete dominios: discriminación y toxicidad, privacidad y seguridad, desinformación, actores malintencionados y uso indebido, interacción humano-computadora, impactos socioeconómicos y ambientales, y seguridad del sistema de IA, fallos y limitaciones como se visualiza en la Figura 15. Además, se estableció que el dominio de la desinformación es el menos estudiado en las revisiones sistemáticas relacionadas con la ética de la IA (Slattery et al., 2024).

Figura 15
Dominios de riesgos éticos de la Inteligencia artificial

Fuente: Elaboración propia.

Segundo, en lo referente al dominio de desinformación, se identificaron como riesgos más frecuentemente analizados en las revisiones de literatura los siguientes: sesgos en los datos (7); desinformación intencionada (6); generación de contenido erróneo (5); generación de contenido falso (5); manipulación de la información (5); manipulación de la opinión pública (5); dificultad en la verificación de las fuentes (4); falta de contexto (4) y el impacto sobre la credibilidad (4) como se visualiza en la Figura 16. Los riesgos con mayor frecuencia de estudio, se asocian al subdominio de información falsa o engañosa, lo cual está asociado a prácticas no intencionales de una cantidad considerable de usuarios de las inteligencias artificiales generativas. Adicionalmente, este tipo de riesgo está altamente relacionado con la ausencia de verificación y validación del conocimiento generado por este tipo de IA.

Figura 16
Subdominios de riesgos de desinformación

Fuente: Elaboración propia.

Tercero, se proponen algunas estrategias de mitigación para los riesgos recurrentes en el uso de la IA en actividades cotidianas en el contexto de la comunicación científica como se visualizan en la Figura 17. En particular, para mitigar el riesgo de alucinaciones, se sugiere configurar en el prompt la variable de configuración temperatura con valores cercanos a 0, lo cual permite controlar el nivel de generación de alucinaciones en los LLM. Además, frente al riesgo de ausencia de verificación y validación del conocimiento generado por la IA generativa (IAG), se recomienda el uso del método de triangulación hermenéutica. Este método cualitativo permite el análisis comparativo del conocimiento generado por la IAG, cotejándolo con los referentes teóricos y la posición del investigador, como una estrategia de verificación. Finalmente, otro riesgo al que se enfrentan los usuarios de la IAG es la falta de contexto. Como estrategia de mitigación, se propone el uso de una estructura de prompt que incluya contexto, indicación, formato y límite.

Figura 17
Formas de mitigar los riesgos de desinformación

Fuente: Elaboración propia.

Finalmente, varias revisiones de literatura sobre los aspectos éticos de la IA han abordado el tema de forma genérica (Carobene et al., 2024; Deng et al., 2023; Hagendorff, 2024). Sin embargo, más recientemente se ha estructurado un repositorio de riesgos con características más integrales y un nivel de desagregación en dominios y subdominios (Slattery et al., 2024). Lo particular de esta investigación es que se realizó una revisión de riesgos de IA focalizada en el dominio de la desinformación, donde se identificaron 73 riesgos asociados a la desinformación en el contexto del uso de la IA.

Abordar los riesgos de la IA en la práctica implica una combinación de estrategias técnicas, organizativas y regulatorias. Específicamente, se enumeran estrategias de gobernanza de IA, mitigación de sesgos, seguridad y privacidad, transparencia y responsabilidad, colaboración, y alfabetización ética y responsable. Estas estrategias ayudan a mitigar los riesgos asociados con la IA y a promover su uso ético, seguro y responsable en diversos ámbitos de la sociedad.

Nuestros resultados muestran la gran diversidad y cantidad de riesgos asociados al dominio de la desinformación, los cuales podrían contribuir al diseño de políticas para un uso ético y responsable de la IA a nivel de instituciones universitarias, nacionales y globales.

El número de riesgos del dominio de desinformación identificados a través de la revisión sistemática de la literatura asciende a 73. Sin embargo, algunos de ellos tienden a confundirse debido a la cercanía conceptual, lo que podría llevar a interpretarlos como si fuesen conceptualmente iguales en un análisis preliminar. Además, existen limitaciones en la interpretación de los riesgos, ya que algunos podrían ser ubicados en otros dominios de riesgos.

CONCLUSIONES Y RECOMENDACIONES

En conclusión, este estudio ha examinado el complejo panorama de los aspectos éticos del uso de la inteligencia artificial, centrando su enfoque en el dominio de la desinformación. Nuestra investigación sobre los desafíos éticos asociados con el uso de la inteligencia artificial en diversos contextos reconoce el valor añadido que las aplicaciones impulsadas por IA pueden aportar. No obstante, también destaca importantes retos éticos derivados tanto del diseño integral de las IA como del uso del conocimiento generado por las IA generativas (IAG).

El uso generalizado de herramientas de IA en diferentes campos académicos y científicos, sumado a la falta o insuficiencia de políticas y normativas claras que regulen el diseño de sistemas de IA y la utilización del conocimiento generado por estas, subraya la necesidad urgente de estudiar y analizar los riesgos asociados con su implementación en la sociedad. Este estudio se centra en el análisis de los riesgos del dominio de desinformación, que incluye los subdominios de información falsa o engañosa y de contaminación del ecosistema informativo y pérdida de la realidad consensuada. En particular, se identificaron 73 riesgos asociados a este dominio, lo que pone en evidencia la amplia diversidad de escenarios en los que la desinformación puede materializarse a través de herramientas de IA.

Estos hallazgos sugieren que, aunque la IA tiene un gran potencial para acelerar el avance científico, su integración efectiva, responsable y ética requiere un desarrollo continuo de directrices y políticas que guíen su uso adecuado. Para ello, es fundamental identificar los riesgos potenciales que podrían materializarse si las herramientas de IA y el conocimiento generado por las IAG no se utilizan de manera apropiada. Dado que el dominio de desinformación es uno de los menos estudiados, esta investigación

proporciona un importante aporte al identificar la diversidad de riesgos involucrados.

Las implicaciones de este estudio van más allá del uso puramente instrumental de las herramientas de IA; subrayan la necesidad de un uso ético y responsable en diversos contextos sociales, lo que plantea nuevos desafíos y requerimientos en la utilización del conocimiento generado por IA.

Las investigaciones futuras deben enfocarse en profundizar el análisis de los riesgos presentes en otros dominios, como la discriminación y toxicidad, privacidad y seguridad, actores maliciosos y uso indebido, interacción humano-computadora, aspectos socioeconómicos y ambientales, así como la seguridad, fallos y limitaciones de los sistemas de IA. Además, es crucial desarrollar estrategias efectivas para mitigar los riesgos identificados en estos dominios.

A medida que la IA continúa transformando los procesos académicos e investigativos en diversos campos, este estudio proporciona una base sólida para promover un uso ético y responsable del conocimiento generado por la IA, optimizando al mismo tiempo las herramientas de IA de uso específico para elevar los estándares de calidad en cada disciplina.

REFERENCIAS BIBLIOGRÁFICAS

Abimbola, C., Eden, C.A., Chisom, O.N., & Adeniyi, I.S. (2024). Integrating AI in education: Opportunities, challenges, and ethical considerations. *Magna Scientia Advanced Research and Reviews, 10*(02), 006–013. https://doi.org/10.30574/msarr.2024.10.2.0039

Al-Rashed, N. (2024). The Impact of Artificial Intelligence on the Future of Media. *International Journal for Scientific Research*., 108 https://doi.org/10.59992/ijsr.2024.v3n7p3

Ballester, V. G., & Vicente, C. T. (2024). El rol de la inteligencia artificial en la publicación científica: perspectivas desde la farmacia hospitalaria. *Farmacia Hospitalaria, 48*(5), 246-251. https://doi.org/10.1016/j.farma.2024.06.002

Carobene, A., Padoan, A., Cabitza, F., Banfi, G., & Plebani, Mo. (2024). Rising adoption of artificial intelligence in scientific publishing: evaluating the role, risks, and ethical implications in paper drafting and review process. *Clinical Chemistry and Laboratory Medicine, 62*(5), pp. 835-843. https://doi.org/10.1515/cclm-2023-1136

Cui, T., Wang, Y., Fu, C., Xiao, Y., Li, S., Deng, X., Liu, Y., Zhang, Q., Qiu, Z., Li, P., Tan, Z., Xiong, J., Kong, X., Wen, Z., Xu, K. & Li, Q. (2024). Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems. *ArXiv*. http://arxiv.org/abs/2401.05778

Critch, A., & Russell, S. (2023). TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. ArXiv*, 1*. https://arxiv.org/abs/2306.06924

Cunha, P.R., & Estima, J. (2023). Navigating the Landscape of AI Ethics and Responsibility. In N. Moniz, Z. Vale, J. Cascalho, C. Silva, R. Sebastião (eds) *Progress in Artificial Intelligence. EPIA 2023. Lecture Notes in Computer Science* (vol. 14115, pp. 92-105). Springer. https://doi.org/10.1007/978-3-031-49008-8_8

Deng, J., Cheng, J., Sun, H., Zhang, Z. & Huang, M. (2023). Towards Safer Generative Language Models: A Survey on Safety Risks, Evaluations, and Improvements*. ArXiv, 1.* http://arxiv.org/abs/2302.09270

Dhawan, S., & Kumar, K. (2024). Ethical Implications of AI in Healthcare. *International Journal of Scientific Research In Engineering And Management, 8*(3), 1-12. https://doi.org/10.55041/IJSREM29006

Electronic Privacy Information Centre. (2023). *Generating Harms: Generative AI's Impact & Paths Forward*. https://epic.org/documents/generating-harms-generative-ais-impact-paths-forward/

Fill, H., Härer, F., Vasic, I., Borcard, D., Reitemeyer, B., Muff, F., Curty, S., & Bühlmann, M. (28-31 de October de 2024). *CMAG: A Framework for Conceptual Model Augmented Generative Artificial Intelligence*. ER2024: Companion Proceedings of the 43rd International Conference on Conceptual Modeling: ER Forum, Pittsburgh, Pennsylvania, USA,

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Stevie Bergman, A., Shelby, R., Marchal, N., Griffin, C., "..." Manyika, J. (2024). The Ethics of Advanced AI Assistants. *ArXiv, 1.* https://doi.org/10.48550/arXiv.2404.16244

Ghani, M., Mustafa, W., Shakir, A., Bakhtiar, D., & Ramadan, G. (2023). *Symbiotic Relationship Between Media Technology and Artificial Intelligence: A Synthesis Analysis*. 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC). Tumkur, India. https://doi.org/10.1109/ICMNWC60182.2023.10435827

Golpayegani, D., Pandit, H.J., & Lewis, D. (2022). AIRO: An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards. *IOS Press, 55,* 51-55. https://doi.org/10.3233/SSW220008

Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. *ArXiv.* http://arxiv.org/abs/2402.08323

Hendrycks, D. & Mazeika, M. (2022). X-Risk Analysis for AI Research. *ArXiv.* https://arxiv.org/abs/2206.05862

Hogenhout, L. (2021). A Framework for Ethical AI at the United Nations. *ArXiv.* http://arxiv.org/abs/2104.12547

Hryciw, B. N., Seely, A. J. & Kyeremanteng, K. (2023). Guiding principles and proposed classification system for the responsible adoption of artificial intelligence in scientific writing in medicine. *Frontiers in Artificial Intelligence, 6,* 1-5. https://doi.org/10.3389/frai.2023.1283353

Infocomm Media Development Authority. (2023). *Cataloguing LLM Evaluations.* https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf

Jiao, J., Afroogh, S., Xu, Y. & Phillips, C. (2024). Navigating LLM Ethics: Advancements, Challenges, and Future Directions. *ArXiv, 1.* https://doi.org/10.48550/arXiv.2406.18841

Khalifa, M. & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update, 5,* 100145. https://doi.org/10.1016/j.cmpbup.2024.100145

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F. & Li, H. (2023). Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *ArXiv*. http://arxiv.org/abs/2308.05374

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W. & Liu, X. (2023). AI vs. Human -- Differentiation Analysis of Scientific Content Generation. *ArXiv.* https://doi.org/10.48550/arXiv.2301.10416

Manoj, K. & Sahil, J. (2025). Ethical issues in the development of artificial intelligence: recognizing the risks. *International Journal of Ethics and Systems, 41*(1), 45–63.https://doi.org/10.1108/IJOES-05-2023-0107

Nah, F. F. H., Zheng, R., Cai, J., Siau, K. & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research, 25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

Niveditha, K. P., Choubey, K., Soni, K., Mishra, P., & Naz, A. (2024). Exploring the Role of Artificial Intelligence in Modern Education. *International Journal of Innovative Research in Computer and Communication Engineering, 12*(5), 8147-8150. https://ijircce.com/admin/main/storage/app/pdf/PeLGRoEovWK8yJozTjh8clpsY2Y0Ae5P0FFoOKOS.pdf

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2024). AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act. *Digital Society, 3*(13), 1-29. https://doi.org/10.1007/s44206-024-00095-1

Peña-Fernández S., Peña-Alonso U., & Eizmendi-Iraola M. (2023). El discurso de los periodistas sobre el impacto de la inteligencia artificial generativa en la desinformación. *Estudios sobre el Mensaje Periodístico, 29*(4), 833-841. https://doi.org/10.5209/esmp.88673

Penabad-Camacho, L., Penabad-Camacho, M. A., Mora-Campos, A., Cerdas-Vega, G., Morales-López, Y., Ulate-Segura, M., Méndez-Solano, A., Nova-Bustos, N., Vega-Solano, M. F. & Castro-Solano, M. M. (2024). Heredia Declaration: Principles on the use of Artificial Intelligence in scientific publishing. *Revista Electrónica Educare, 28*(S), 1-10. https://doi.org/10.15359/ree.28-S.19967

Rodrigues, F., Newell, R., Babu, G. R., Chatterjee, T., Sandhu, N. K., & Gupta, L. (2024). The social media Infodemic of health-related misinformation and technical solutions. *Health Policy and Technology, 13*(2), 100846. https://doi.org/10.1016/j.hlpt.2024.100846

Rogers, W.A., Draper, H., & Carter, S.M. (2021). Evaluation of artificial intelligence clinical applications: Detailed case analyses show value of healthcare ethics approach in identifying patient care issues. *Bioethics, 35*(7), 623-633. https://doi.org/10.1111/bioe.12885

Samuel-Okon, A. D. (2024). Smart Media or Biased Media: The Impacts and Challenges of AI and Big Data on the Media Industry. *Asian Journal of Research in Computer Science, 7*(7), 128–144. https://doi.org/10.9734/ajrcos/2024/v17i7484

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N. 'mah, Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 723-741). Association for Computing Machinery. https://doi.org/10.1145/3600211.3604673

Slattery, P., Saeri, A., Grundy, E.A., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S. & Thompson, N. (2024). The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. *ArXiv, 1.* https://doi.org/10.48550/arXiv.2408.12622

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., ... & Wright, D. (2023). A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review, 56*(11), 12799-12831. https://doi.org/10.1007/s10462-023-10420-8

Sun, H., Zhang, Z., Deng, J., Cheng, J., & Huang, M. (2023). Safety Assessment of Chinese Large Language Models. *ArXiv,* https://arxiv.org/abs/2304.10436

Ustinovich, E. (2024). Generative artificial intelligence in the electoral processes of 2024 in the world: disinformation campaigns and online trolls. *Social''naja Politika I Social''noe Partnerstvo (Social Policy and Social Partnership, (3)gener.* https://doi.org/10.33920/pol-01-2403-03

Vidgen, B., Agrawal, A., Ahmed, A. M., Akinwande, V., Al-Nuaimi, N., Alfaraj, N., Alhajjar, E., Aroyo, L., Bavalatti, T., Blili-Hamelin, B., Bollacker, K., Bomassani, R., Boston, M.F., Campos, S., Chakra, K., Chen, C., Coleman, C., Coudert, Z. D., Derczynski, L., "..." Vanschoren, J. (2024). Introducing v0.5 of the AI Safety Benchmark from MLCommons. *ArXiv.* http://arxiv.org/abs/2404.12241

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V. & Isaac, W. (2023). Sociotechnical Safety Evaluation of Generative AI Systems. *ArXiv.* http://arxiv.org/abs/2310.11986

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J, Legassick, S., Irving, G. & Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* 214–229. https://doi.org/10.1145/3531146.3533088

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A. & Gabriel, I. (2021). Ethical and social risks of harm from Language Models. *ArXiv.* http://arxiv.org/abs/2112.04359

Williamson, S. M., Prybutok, V. (2024). Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare. *Applied Sciences, 14*(2), 675. https://doi.org/10.3390/app14020675

Zeng, Y., Klyman, K., Zhou, A., Yang, Y., Pan, M., Jia, R., Song, D., Liang, P., & Li, B. (2024). AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. *ArXiv,* https://arxiv.org/abs/2406.17864

Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J. & Huang, M. (2023). SafetyBench: Evaluating the safety of Large Language Models. *ArXiv*. https://doi.org/10.48550/arXiv.2309.07045

AUTORES

William Mauricio Rojas Contreras

Postdoctorado en Investigación social para la innovación. Doctor en Educación, Magister en Ciencias Computacionales, Especialista en Ingeniería del software, Ingeniero de Sistemas. Ha publicado artículos en revistas nacionales e internacionales y autor de libros en el área de Ciencias Computacionales. Líneas de trabajo Ingeniería del Software, inteligencia artificial y Ciencias de la Computación. Actualmente es profesor titular de tiempo completo programa de Ingeniería de Sistemas de la Universidad de Pamplona. Investigador Senior MinCiencias y Director del grupo de Investigación de Ciencias Computacionales – CICOM Categoría A MinCiencias.

https://orcid.org/0000-0002-9055-5792

Olga Liliana Rojas Contreras

Liliana Rojas Contreras, Doctora en Ciencia de los Alimentos por la Universidad Autónoma de Barcelona (España), Magíster en Ciencia y Tecnología de Alimentos por la Universidad de Pamplona y Especialista en Química Ambiental por la Universidad Industrial de Santander, con sólida formación en Microbiología. Actualmente, es docente en la Universidad de Pamplona y su investigación se centra en microbiología de alimentos, micotoxinas y microbiología ambiental. Es miembro activo del Grupo de Investigación en Microbiología y Biotecnología (GIMBIO) y coordinadora del Semillero en Microbiología y Biotecnología (SIMBIO). Ha desempeñado roles de dirección en el departamento de Microbiología, con destacada trayectoria en formación y evaluación académica. Ha publicado numerosos trabajos científicos y ha contribuido significativamente a la investigación aplicada en alimentación y medio ambiente. Su experiencia abarca también la administración universitaria y el

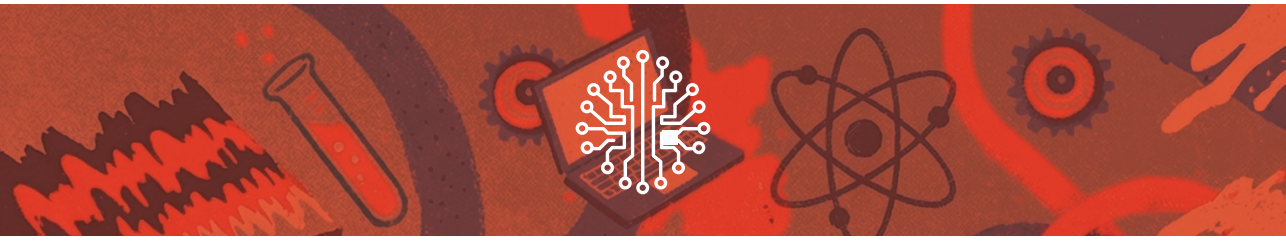
liderazgo en proyectos de investigación enfocados en microbiología y micotoxicología, consolidándola como una profesional reconocida en su campo.

https://orcid.org/0000-0001-9184-9031

Eliana Caterine Mojica Acevedo

Comunicadora Social – periodista y organizacional, magister en Gerencia Educativa. Docente investigadora y consultora en empresas públicas y privadas en el fortalecimiento de procesos organizacionales. Docente Titular de la Universidad de Pamplona, representante de la Facultad de Artes y Humanidades ante Comité de Investigaciones de la Universidad. Actualmente, Jefe Sello Editorial de la Universidad de Pamplona.

https://orcid.org/0000-0002-7659-3370



SELLO EDITORIAL
UNIVERSIDAD DE PAMPLONA

ISBN (Digital): 978-628-7656-87-1