IPIE

International Panel on the Information Environment

Respuesta a la desinformación generada por inteligencia artificial generativa

Resultados de un metaanálisis de la evidencia científica

Resumen para responsables de las políticas
2026.2

Número DOI: 10.61452/QXAF2136

A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-Lopez, P. N. Howard

Respuesta a la desinformación generada por inteligencia artificial generativa

Resultados de un metaanálisis de la evidencia científica

Resumen para responsables de las

Cómo citar:
International Panel on the Information Environment [A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-López, P. N. Howard (eds.)], “Responding to Generative AI Misinformation: Results from a Meta-Analysis of Scientific Evidence,” Zurich, Switzerland: IPIE, 2026. Resumen para responsables de políticas, SFP2026.2, doi: 10.61452/HFLQ3407.

SINOPSIS

La inteligencia artificial generativa (IAGen) puede producir ahora rápidamente grandes volúmenes de texto, imágenes, audio y vídeo engañosos con herramientas de fácil acceso. Esto resulta importante para los responsables de políticas, porque el contenido falso o engañoso puede adaptarse a audiencias específicas. Ese contenido puede propagarse con rapidez e imitar comunicaciones auténticas, lo que dificulta que las personas determinen qué es verdadero y qué no.

Este Resumen para responsables de políticas sintetiza las principales conclusiones del Informe de síntesis del IPIE (SR2026.2) sobre los efectos de la desinformación producida con IAGen y las medidas con mayor probabilidad de reducir su influencia.

La evaluación se basa en un metaanálisis a gran escala de evidencia científica experimental. Se basa en 60 estimaciones del efecto procedentes de ensayos controlados aleatorizados, extraídas de 24 publicaciones revisadas por pares en las que participaron 33 801 personas, publicadas entre 2018 y 2025.

El informe llega a cuatro conclusiones principales:

1. La desinformación textual generada por IAGen plantea actualmente mayores riesgos de persuasión que la desinformación visual.
2. La base de evidencia actual deja fuera a la mayor parte del mundo. La investigación se concentra en países de renta alta y de habla inglesa, lo que deja importantes lagunas de conocimiento.
3. La intervención con una eficacia más constante consiste en proporcionar a los usuarios información correctiva de forma preventiva, para que puedan evaluar por sí mismos la exactitud y la credibilidad de la misma.
4. El etiquetado es eficaz cuando se aplica de forma coherente. El etiquetado de contenidos suele reducir la credibilidad percibida, pero su impacto varía considerablemente en función de cómo las empresas crean y aplican las etiquetas.

Las implicaciones para las políticas son evidentes. Los responsables de formular políticas deben centrarse prioritariamente en combatir la desinformación textual generada por IAGen y respaldar la información preventiva y correctiva como eje fundamental de actuación. El etiquetado debe considerarse una medida que exige un diseño cuidadoso y pruebas previas rigurosas. Es necesario ampliar la investigación independiente mucho más allá de los países de habla inglesa y de renta alta.

El acceso de los investigadores a los datos de las plataformas y de los modelos es esencial para que la sociedad comprenda mejor la desinformación generada por IAGen y para comprobar qué medidas de protección funcionan realmente en la práctica.

INTRODUCCIÓN

Los sistemas de inteligencia artificial generativa (IAGen) pueden producir texto, imágenes, audio y vídeos que, en ocasiones, resultan difíciles de distinguir de los contenidos creados por personas. Los contenidos generados por IA se han generalizado y ya aparecen en procesos electorales de todo el mundo, a menudo procedentes de fuentes desconocidas o malintencionadas, además de influir cada vez más en la percepción pública en ámbitos como la salud pública, los conflictos y los mercados, con importantes efectos en el mundo real [1]. Estos sistemas incluyen grandes modelos de lenguaje que generan texto, herramientas de síntesis de imagen y audio y sistemas de creación de vídeo. También pueden generar contenidos engañosos o perjudiciales, como noticias falsas, rumores, propaganda o deepfakes. En este informe, todos estos fenómenos se engloban bajo el término «desinformación».

Este Resumen para responsables de políticas presenta las principales conclusiones de un metaanálisis de la evidencia experimental sobre los efectos que tiene en las personas la exposición a desinformación creada con IAGen [2]. También examina dos medidas para contrarrestarla que cuentan con evidencia experimental suficiente para el debate de políticas: la información correctiva y el etiquetado de contenidos. El informe pone el foco en los ensayos controlados aleatorizados porque estos estudios permiten realizar comparaciones más fiables entre experimentos y calcular de forma más sistemática los efectos medios.

La evidencia apunta a un entorno informativo que evoluciona con gran rapidez. El informe constata que la respuesta del público ante la desinformación textual y visual está empezando a diferenciarse. También señala que la base de evidencia sigue siendo limitada en cuanto a su alcance geográfico y sus métodos. Al mismo tiempo, identifica dos respuestas prácticas capaces de reducir la credibilidad percibida de la desinformación generada por IAGen, aunque una de ellas ofrece resultados más constantes que la otra.

Para construir esta base de evidencia, la revisión consultó dos grandes bases de datos bibliográficas, Scopus y Web of Science, incorporó recomendaciones de expertos, fusionó los resultados, eliminó duplicados y examinó un conjunto final de 6952 publicaciones. Tras el cribado de elegibilidad y la codificación del texto completo, 87 publicaciones cumplieron los criterios para una revisión detallada. Los controles adicionales de sesgo, el cribado y las revisiones del protocolo siguiendo las recomendaciones de Cochrane dieron lugar a un subconjunto de 24 estudios especialmente relevantes para su inclusión en el metaanálisis. El Informe de síntesis principal utiliza un metaanálisis de efectos aleatorios de tres niveles y evalúa el riesgo de sesgo mediante la «herramienta revisada de riesgo de sesgo en ensayos aleatorizados». Para consultar todos los detalles metodológicos, véase el Informe de síntesis completo [2].

RESULTADO 1. EL TEXTO ENGAÑOSO GENERADO POR IAGEN ES LA PRINCIPAL AMENAZA

CONCLUSIÓN CLAVE

La desinformación textual generada por IA generativa plantea actualmente mayores riesgos de persuasión que la desinformación visual.

El informe muestra una diferencia clara en la percepción de fiabilidad entre el contenido textual y el contenido visual generado por IAGen. Los estudios más recientes indican que el texto falso generado por IAGen suele percibirse como más exacto, creíble o verosímil que la información verdadera. En cambio, la información visual falsa generada por IAGen, como los deepfakes, suele percibirse como menos creíble que la información verdadera. En los estudios visuales realizados después de 2021, la estimación combinada apunta a una disminución moderada de la credibilidad percibida.

Esto no significa que la desinformación visual sea inocua o carezca de importancia. Los deepfakes siguen entrañando riesgos, y el informe señala que la respuesta del público podría volver a cambiar a medida que mejoren los sistemas visuales. No obstante, la evidencia disponible indica que la desinformación textual merece una atención más urgente por parte de las políticas públicas. Esto es importante porque los sistemas centrados en texto pueden ser económicos, escalables, fáciles de personalizar y estar cada vez más integrados en buscadores, servicios de mensajería e interfaces conversacionales. El informe también señala que los resultados generados por la IA conversacional pueden ser mucho más persuasivos que los mensajes estáticos generados por IA, lo que sugiere que las estimaciones actuales podrían ser prudentes.

El patrón general de los resultados es lo bastante sólido como para respaldar un mensaje claro para las políticas públicas: en este momento, el riesgo persuasivo más inmediato procede del texto falso o engañoso, no solo de los deepfakes visuales.

RESULTADO 2: LAGUNAS DE EVIDENCIA A ESCALA MUNDIAL

El informe subraya claramente que la mayoría de las publicaciones examinadas se centran en países de habla inglesa y de renta alta, como Austria, Alemania, los Países Bajos, Corea del Sur, el Reino Unido y los Estados Unidos. La evidencia disponible se limita a artículos completos en inglés, y cinco publicaciones examinadas en otros idiomas no cumplieron los criterios de elegibilidad. En consecuencia, el conocimiento actual sobre la desinformación generada por IAGen se concentra en un número reducido de países, lenguas y entornos mediáticos.

Se trata de una limitación importante. Indica que los responsables de formular políticas, los organismos reguladores, los investigadores y las empresas siguen disponiendo de un conocimiento limitado sobre cómo opera la desinformación generada por IAGen en muchas otras regiones. El informe afirma explícitamente que esto crea una importante

CONCLUSIÓN CLAVE

La base de evidencia actual deja fuera de su alcance a la mayor parte del mundo, tanto países como lenguas.

laguna de conocimiento más allá de los pocos países y lenguas que publican sobre este tema. Esta laguna es relevante para las sociedades multilingües, la regulación transfronteriza y los esfuerzos por desarrollar salvaguardias eficaces en distintas lenguas, culturas y sistemas políticos.

Además, la evidencia se ve limitada por la rapidez con la que avanza la tecnología. Muchos experimentos analizan sistemas antiguos, como GPT-2, o herramientas iniciales como FaceSwap. El informe advierte de que la investigación científica suele ir por detrás de la tecnología disponible en cada momento. Como resultado, algunas políticas se basan en evidencia obtenida con sistemas menos avanzados que los que ya utiliza el público. Por ello, resulta imprescindible impulsar investigación independiente en contextos más amplios, a fin de que la regulación y la gobernanza se mantengan alineadas con la tecnología que está configurando el debate público.

RESULTADO 3. LAS CORRECCIONES PREVENTIVAS PUEDEN AYUDAR

La información correctiva preventiva incluye breves advertencias educativas, recordatorios o avisos, así como otras formas de preparar a los usuarios antes de que se expongan a contenidos potencialmente engañosos. Por ejemplo, algunas intervenciones ofrecen explicaciones breves sobre cómo los deepfakes permiten manipular vídeos de forma realista o recuerdan a los usuarios que los sistemas de IAGen pueden cometer errores. Estos enfoques buscan mejorar la capacidad de los usuarios para valorar si un contenido es fiable.

En los estudios posteriores a 2020, la estimación combinada muestra que la información correctiva preventiva reduce la percepción de exactitud, credibilidad o aceptación de la desinformación generada por IAGen, con un tamaño del efecto de pequeño a moderado. Estos resultados son relativamente coherentes en los estudios revisados: la variación entre sus resultados es baja y los efectos se observan en distintas poblaciones y diseños experimentales dentro de este subconjunto de evidencia.

CONCLUSIÓN CLAVE

La información correctiva preventiva es la intervención con una eficacia más constante.

Estos resultados indican que este tipo de intervenciones probablemente seguirá siendo eficaz a corto plazo,, aunque la mayor parte de la evidencia disponible procede de estudios sobre generaciones anteriores de sistemas de IAGen.

La Revisión de síntesis también destaca una salvedad importante: el efecto positivo de la información correctiva preventiva se observa cuando se pide a las personas que evalúen la exactitud o la credibilidad del contenido. La evidencia es menos uniforme cuando los estudios se centran únicamente en la detección. Investigaciones anteriores sobre deepfakes concluyeron que las advertencias podían, en ocasiones, aumentar la desconfianza tanto hacia los vídeos reales como hacia los manipulados. Esto coincide con una preocupación más amplia en la investigación y las políticas sobre desinformación: las narrativas falsas y la propaganda suelen buscar que

las personas desconfíen de toda la información pública. Por tanto, las intervenciones deben diseñarse para mejorar la capacidad de las personas de distinguir entre contenidos fiables y contenidos engañosos.

La información correctiva preventiva es una de las medidas más estudiadas que pueden aplicarse en línea con relativa rapidez, especialmente en las plataformas de redes sociales. Algunos ejemplos son las notas que advierten de que la IAGen puede producir errores, las explicaciones sobre cómo los deepfakes permiten manipular vídeos de forma realista y los materiales educativos breves sobre fenómenos como las alucinaciones de la IA. Según la disciplina, estas medidas reciben nombres como asesoramiento, inoculación, desacreditación preventiva, activación previa o etiquetado de advertencia.

RESULTADO 4. EL ETIQUETADO FUNCIONA CUANDO ES COHERENTE.

CONCLUSIÓN CLAVE

El etiquetado de contenidos sigue siendo prometedor, pero solo si es claro y coherente.

El etiquetado de contenidos consiste en añadir a la información etiquetas breves visuales, textuales o multimodales. Estas etiquetas indican que la información puede haber sido creada con IAGen, haber sido alterada o ser engañosa. En el análisis combinado, las etiquetas se asocian con una reducción pequeña, pero estadísticamente significativa, de la credibilidad percibida de la información engañosa generada por IAGen. Es un resultado alentador, pero no se observa de forma uniforme en todos los casos. El informe detecta una variación considerablemente mayor en la evidencia sobre el etiquetado que en la relativa a la información correctiva preventiva.

Debido a que la evidencia es limitada, los efectos varían de forma considerable y algunos estudios no encuentran efecto alguno. Las etiquetas pueden reducir la credibilidad en algunas situaciones, pero no en otras. Esta evaluación destaca varios factores que probablemente moderan su efecto, como el diseño, la formulación y la presentación de la etiqueta, el tipo de contenido, el contexto experimental y la fuente de la etiqueta. En otras palabras, la forma en que se diseñan y aplican las etiquetas importa al menos tanto como el hecho de utilizarlas.

La base de evidencia sobre el etiquetado también es limitada. Los estudios de esta muestra, relativamente pequeña, se centran casi por completo en participantes de Estados Unidos. Esto dificulta saber cómo funcionarán las etiquetas en distintas lenguas, culturas o sistemas jurídicos. Por tanto, la conclusión adecuada no es que el etiquetado no funcione. Más bien, el etiquetado puede ser útil, pero solo si las empresas cuidan el diseño de estas etiquetas y las prueban en diversos contextos, en lugar de dar por supuesto que funcionan por defecto en todas partes.

CONCLUSIÓN

Este Resumen para responsables de políticas pone de relieve un patrón de riesgo cambiante. La desinformación creada mediante IAGen textual supone una amenaza persuasiva mayor que la generada mediante IAGen visual. Sin embargo, esto no elimina los peligros de los deepfakes ni de otros medios sintéticos. Lo que indica es que los responsables de formular políticas no deben permitir que la notoriedad del engaño visual desvíe la atención del creciente poder persuasivo del texto falso generado por IA.

Este informe retoma los resultados de evaluaciones anteriores del IPIE [3], [4] y propone dos respuestas prácticas. La información correctiva preventiva es la intervención más sólida y con respaldo más consistente en la base de evidencia actual, especialmente cuando se ofrece antes de la exposición y ayuda a mejorar la capacidad de evaluación. El etiquetado de contenidos también puede ser útil, pero su eficacia varía y depende del contexto. Las etiquetas funcionan mejor cuando se diseñan con cuidado y se aplican de forma clara y coherente.

El último punto se refiere a la propia evidencia. La mayor parte del mundo sigue quedando fuera del alcance de la investigación actual. Esto limita la capacidad de los responsables de políticas, las plataformas y los desarrolladores de modelos de IA de propósito general para entender cómo se propagan estos riesgos en distintas lenguas, culturas y sistemas mediáticos. Por tanto, son esenciales la investigación independiente, la síntesis continua de la evidencia y un mejor acceso de los investigadores a los datos de las plataformas y los modelos. Sin todo ello, las políticas públicas seguirán siendo reactivas, incompletas y demasiado estrechas de miras en un entorno informativo que evoluciona con rapidez.

REFERENCIAS

[1] International Panel on the Information Environment [I. Trauthig, P. N. Howard, S. Valenzuela (eds.)], "The Role of Generative AI Use in 2024 Elections Worldwide," Zurich, Switzerland: IPIE, 2025. Technical Paper, TP2025.2, doi: 10.61452/HZUE9853.

[2] International Panel on the Information Environment [A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-López, P. N. Howard (eds.)], "Effects of Misinformation Produced with Generative AI and its Countermeasures: A Meta-Analysis of Experimental Scientific Evidence," Zurich, Switzerland: IPIE, 2026. Synthesis Report, SR2026.2, doi: 10.61452/UGTR3022.

[3] International Panel on the Information Environment, "Countermeasures for Mitigating Digital Misinformation: A Systematic Review," IPIE, Zurich, Switzerland, SR2023.1, July 2023. [Online]. Disponible en: https://www.ipie.info/research/sr2023-1

[4] International Panel on the Information Environment, "Platform Responses to Misinformation: A Meta-Analysis of Data," IPIE, Zurich, Switzerland, SR2023.2, July 2023. [Online]. Disponible en: https://www.ipie.info/research/sr2023-2

AGRADECIMIENTOS

Colaboradores

Redactores: Aliaksandr Herasimenka (científico consultor, Reino Unido), Sebastián Valenzuela (director científico del IPIE y presidente del Comité de Ciencia y Metodología, Chile), Shelley Boulianne (miembro del Comité de Ciencia y Metodología del IPIE, Canadá), Frank Esser (miembro del Comité de Ciencia y Metodología del IPIE, Suiza), Lisa M. Given (miembro del Comité de Ciencia y Metodología del IPIE, Canadá/Australia), Stephan Lewandowsky (miembro del Comité de Ciencia y Metodología del IPIE, Australia/Reino Unido), Eva M. Navarro-López (miembro del Comité de Ciencia y Metodología del IPIE, España/Reino Unido/México), Philip Howard (presidente y director ejecutivo del IPIE, Canadá/Reino Unido). Ayudantes de investigación: Anna George y Xianlingchen Wang. Revisiones generales independientes: George Georgarakis y Mathias Harrer. Diseño: Domenico Di Donna. Corrección: Beverley Sykes. Agradecemos el apoyo de la secretaría del IPIE: Lola Gimferrer, Jessica Gold, Wiktoria Schulz, Donna Seymour, Anna Staender y Alex Young.

Financiadores

El International Panel on the Information Environment (IPIE) agradece el apoyo de sus financiadores. Para consultar la lista completa de socios financiadores, consultewww.ipie.info. Las opiniones, hallazgos, conclusiones o recomendaciones de este informe corresponden al IPIE y no necesariamente reflejan los puntos de vista de sus patrocinadores.

Declaración de intereses

Los informes del IPIE son elaborados y revisados por una red global de investigadores asociados que forman paneles científicos especializados. Todos los colaboradores y los revisores realizan sus declaraciones de intereses, las cuales son revisadas por el IPIE en las correspondientes etapas de trabajo.

Cita recomendada

Un *Resumen para responsables de la formulación de políticas del IPIE* ofrece una síntesis de alto nivel del estado del conocimiento y está dirigido a un público amplio. Los *Informes de síntesis* del IPIE hacen uso de técnicas de metaanálisis, exámenes sistemáticos y otras herramientas para la agrupación de pruebas, la generalización de conocimientos y la formación de consenso científico, y están escritos para un público especializado. Un *Documento técnico del IPIE* aborda cuestiones metodológicas concretas o proporciona un análisis de políticas sobre un problema regulatorio específico. Todos los informes están disponibles en la página web del IPIE (www.IPIE.info).

Este documento debe citarse de la siguiente manera:

International Panel on the Information Environment [A. Herasimenka, S. Valenzuela, S. Boulianne, F. Esser, L. M. Given, S. Lewandowsky, E. M. Navarro-López, P. N. Howard (eds.)], "Responding to Generative AI Misinformation: Results from a Scientific Meta-Analysis," Zurich, Switzerland: IPIE, 2026. Resumen para responsables de políticas, SFP2026.2, doi: 10.61452/HFLQ3407.

Información sobre los derechos de autor

SOBRE EL IPIE

El Panel Internacional sobre el Entorno de Información (IPIE) es una organización científica independiente y global dedicada a facilitar el conocimiento científico más práctico posible sobre las amenazas para el entorno mundial de información. Con sede en Suiza, la misión del IPIE es ofrecer a los responsables de políticas, a la industria y a la sociedad civil evaluaciones científicas independientes sobre el entorno global de la información mediante la organización, la evaluación y la promoción de la investigación, con el objetivo general de mejorar el entorno global de la información. Cientos de investigadores de todo el mundo colaboran en los informes del IPIE.

Para más información, póngase en contacto con el Panel Internacional sobre el Entorno de Información (IPIE), secretariat@IPIE.info. Seefeldstrasse 123, P.O. Box, 8034 Zurich, Suiza.

International Panel on
the Information
Environment

Seefeldstrasse 123
P.O. Box 8034 Zurich
Suiza