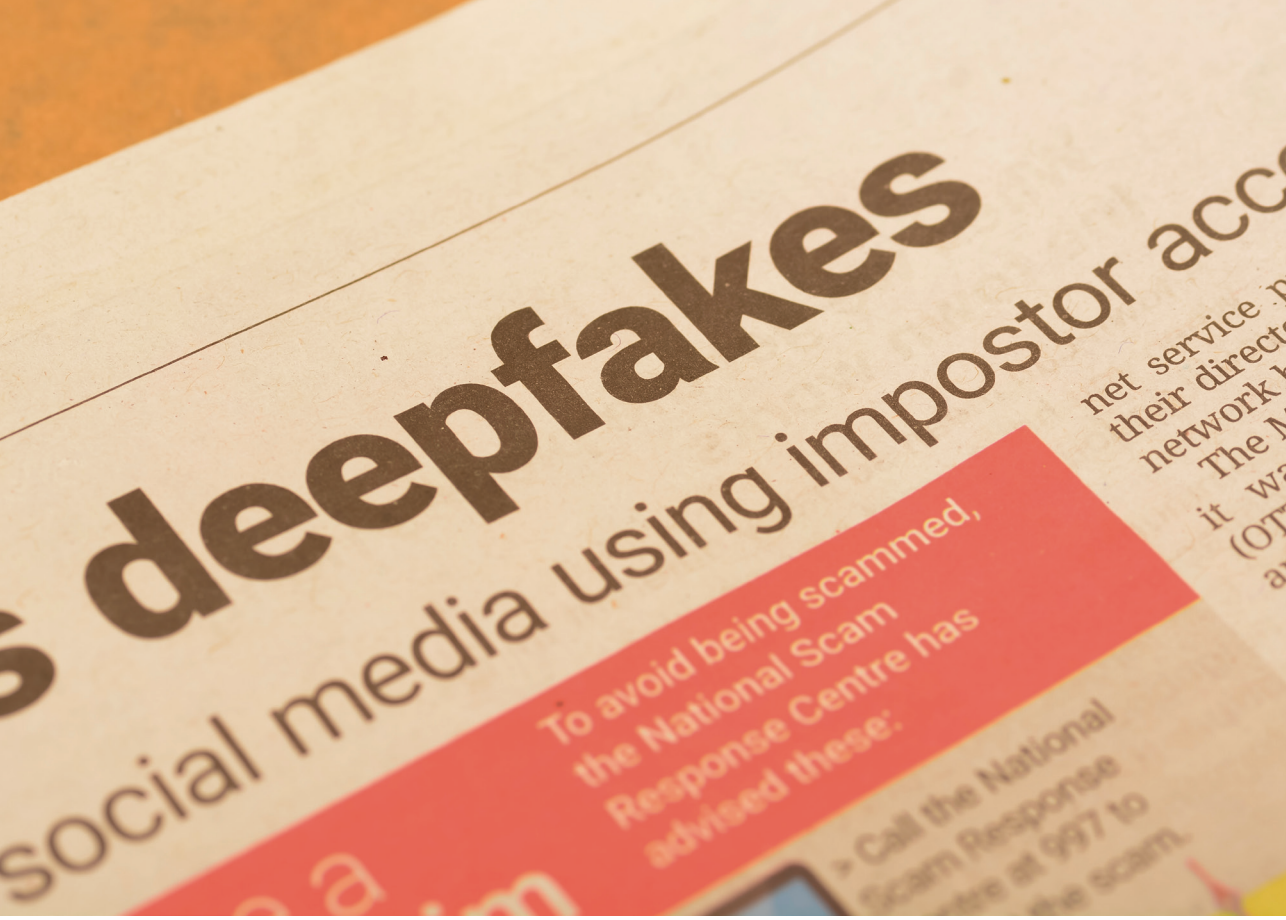


Personalizar las *fakes*: ¿hacia el apocalipsis de la desinformación?

Christoph Nehring

Este artículo explora los efectos y las implicaciones de la desinformación generada por la IA. Examina sus formas, en particular los *deepfakes,* y su impacto en las elecciones recientes y futuras. También ofrece ideas prácticas para identificar y combatir la desinformación generada por la IA, con especial atención al papel de los *influencers,* los periodistas y otros profesionales de los medios, y a los desafíos únicos que enfrentan.

La inteligencia artificial está transformando el panorama informativo global, creando oportunidades y riesgos sin precedentes. La generación, difusión y amplificación de la desinformación y la misinformación (información errónea) son ejemplos destacados. A pesar del miedo y la confusión generalizados, el conocimiento empírico sobre la desinformación generada por IA, sus formas, impacto y efectos sigue siendo limitado, lo que alimenta la incertidumbre, el temor, la desconfianza y la demanda de información equilibrada y de calidad.

Desinformación, *deepfakes* y manipulación

Ya en 2023 los expertos en desinformación señalaban el potencial de la IA generativa como un arma de engaño masivo, que potencia y amplifica la desinformación y la misinformación. Aunque estos escenarios catastrofistas aún no se han materializado, la IA tiene varias cualidades que afectan significativamente la producción y distribución de desinformación y misinformación. La IA puede hacer que la desinformación sea:

- *más rápida,* tanto en la creación de contenidos como en su distribución automática;
- *más barata,* p. ej., automatizando la producción y la distribución, reduciendo los recursos humanos y financieros;
- *más persuasiva,* p. ej., utilizando *deepfakes* superrealistas;
- *más personalizada,* p. ej., utilizando IA para analizar datos y ajustar los mensajes para alcanzar a públicos específicos;
- *de mayor alcance,* p. ej., usando bots de IA y automatizando la distribución de desinformación, o simplemente porque las herramientas de IA están disponibles para todos los usuarios de redes sociales.

Experimentos de hackers y periodistas muestran en un ejemplo que el costo de usar ChatGPT para operar una web de noticias falsas totalmente automatizada bajó de USD 400 a USD 105 entre 2023 y 2024.

Formas de desinformación de las IA

La revolución de la GenAI impacta en el software que crea contenidos, textos, imágenes, videos y audio. Las formas conocidas de desinformación generada por IA incluyen:

a. *Sitios web de noticias falsas.* Aunque difíciles de detectar, se han identificado miles de sitios cuyos contenidos (texto, imágenes, videos) son generados por ChatGPT u otros chatbots. Algunos, como Electionwatch o TheDCWeekly, se enfocan en la desinformación organizada sobre la política estadounidense y las elecciones presidenciales de 2024, mientras que otros son sitios comerciales que reescriben y publican noticias antiguas con fines lucrativos.

b. *Imágenes de* IA. Las imágenes generadas por IA están inundando redes sociales, servicios de mensajería y sitios web. Algunos muestran a personas, generalmente políticos, en situaciones ficticias (como Donald Trump bailando con menores) o describen eventos que nunca ocurrieron (como un atentado terrorista al Pentágono). Si bien algunas páginas profesionales de noticias falsas, probablemente respaldadas por actores estatales, usan imágenes generadas por IA junto con artículos falsos, la mayoría de estas imágenes son creadas y compartidas por usuarios comunes en redes sociales y foros. Este tipo de imágenes se difunden especialmente durante el conflicto en Oriente Próximo, destacando los daños de la guerra y las víctimas en Gaza. En algunos casos, las imágenes de IA difundidas en redes sociales llegaron a bases de datos de stock (p. ej., Adobe Stock) donde se vendieron para uso comercial. Por otro lado, los usuarios ucranianos recurren cada vez más a imágenes generadas por IA para expresar apoyo al ejército ucraniano en la guerra contra Rusia. Esta tendencia ilustra los efectos de la democratización de las herramientas genai y su uso indebido.

c. *Deepfakes.* Los llamados *deepfakes* (derivados de *deep learning* y *fakes*) son contenidos de video y audio producidos o manipulados por IA. Existen varios tipos de *deepfakes*, que varían según su aplicación (p. ej., el intercambio de caras en la pornografía o en llamadas fraudulentas) o la intención detrás de ella. Los *deepfakes* creados para desinformación política han aparecido en diversos contextos, como la guerra rusa contra Ucrania y, en particular, durante las campañas electorales en todo el mundo (ver más abajo). En la mayoría de los casos se utilizan para crear pruebas falsas que desacreditan declaraciones o posturas, o para representar participación en actos ilegales o pornografía. Sus víctimas suelen ser personas públicas, como celebridades, políticos, CEO, *influencers* y periodistas. Los *deepfakes* generan un alto nivel de temor y confusión en el público debido a: a) la impresionante calidad de las falsificaciones; b) su capacidad para convencer y persuadir el público; c) la falta de software y métodos de detección fiables; y d) la inseguridad e incapacidad del público para reconocer y enfrentar estos contenidos. En las siguientes secciones de este ensayo se abordarán los *deepfakes* como uno de los ejemplos más urgentes y peligrosos de desinformación generada por IA.

Expertos e investigadores estatales han encontrado pruebas empíricas de la existencia de todas estas formas de desinformación generada por IA. Sin embargo, el llamado *detection challenge* 'reto en la detección' del

contenido de IA dificulta evaluar y determinar con precisión la cantidad y el alcance real de la desinformación generada por IA. Hasta la fecha no existe un método de detección cien por ciento preciso para contenidos generados por IA, ni filtros automáticos de carga ni servicios de eliminación, etc. Esto significa que, aunque la calidad y cantidad de la desinformación generada por IA están aumentando rápidamente, su alcance real sigue siendo difícil de evaluar.

Deepfakes y elecciones en 2023 y 2024

En los últimos dos años, la desinformación de IA, especialmente los *deepfakes*, se ha convertido en un arma para influir en campañas políticas y elecciones. En la mayoría de los casos, la tecnología se ha usado para crear videos o audios de políticos, candidatos, periodistas y otras figuras públicas en situaciones negativas y desacreditadoras. Algunos atacan la reputación de individuos para socavar su credibilidad, imagen y reputación pública, mientras que otros son parte de campañas políticas negativas que buscan desacreditar opiniones, decisiones o eventos políticos. Todos ellos intentan influir en el comportamiento de los votantes, difundiendo deliberadamente información falsa, engañosa o descontextualizada, creada artificialmente.

En otros casos, los *deepfakes* se usan en campañas políticas oficiales. Estos *deepfakes* se distinguen porque: a) están vinculados a una fuente *oficial* como un candidato, partido, institución u organización; b) están (frecuentemente) etiquetados como contenido generado por IA; y c) no contienen necesariamente información falsa. Durante las elecciones al Parlamento Europeo de junio de 2024, varios partidos de extrema derecha y derecha (p. ej., Francia e Italia) usaron tecnología para promover sus mensajes y narrativas mediante memes, imágenes y canciones generadas por IA. En *Pakistán*, el ex primer ministro Imran Khan y su equipo usaron *deepfakes* para incluirlo en videos de campaña mientras estaba encarcelado. En *India, Indonesia y Filipinas,* partidos y equipos de campañas crearon *deepfakes* de políticos fallecidos o figuras públicas populares para sus campañas electorales. Durante las elecciones presidenciales en *Argentina,* ambos candidatos y sus equipos utilizaron extensamente todas las formas de IA generativa (imágenes, videos, texto) en sus campañas. Esto incluyó también videos malintencionados de ambos candidatos, que cruzaron la línea entre la campaña y la desinformación al difundir deliberadamente mentiras provocadoras. En *México,* la entonces candidata presidencial y ex jefa de Gobierno de Ciudad de México, Claudia Sheinbaum, apareció en un video que supuestamente promovía un esquema financiero fraudulento, lo que dañó su credibilidad política. En todos los países y todas las elecciones de 2024, se vieron *deepfakes* políticos diseñados para desacreditar a candidatos y promover narrativas, mayormente agresivas. Durante las elecciones presidenciales de Estados Unidos, el uso de IA generativa fue especialmente prominente. Ambas partes —canales oficiales y simpatizantes— publicaron imágenes generadas por IA para transmitir sus

» La desinformación de IA, especialmente los *deepfakes*, se ha convertido en un arma para influir en campañas políticas y elecciones. «

mensajes. Sin embargo, en redes sociales circularon formas más peligrosas de *deepfakes*, como *robollamadas* con la voz del presidente Joe Biden instando a no votar, *AI-fakes* mostrando a Taylor Swift apoyando a Donald Trump, y contenidos falsos creados por IA que supuestamente aparecían en el libro de J. D. Vance.

A pesar de que los *deepfakes* estuvieron presentes en todas las elecciones de 2024, las pruebas empíricas sugieren que, contrariamente a las expectativas apocalípticas, no tuvieron un impacto significativo en los resultados. Hasta ahora, solo en dos casos los *deepfakes* ocurridos en las 48 horas previas a las elecciones han tenido una influencia decisiva. En Eslovaquia, un de audio en el que un candidato presuntamente discutía la compra de votos de minorías pareció afectar directamente los resultados, aunque no inclinó los resultados finales a favor de otro candidato. Por otra parte, durante las elecciones presidenciales en Turquía, un video *deepfake* de contenido pornográfico que involucraba a uno de los candidatos resultó en su retirada de la contienda electoral. Obviamente, esto influyó en el resultado de las elecciones, pero dado que todas las encuestas ya apuntaban al presidente en funciones como claro vencedor, el *deepfake* puede haber tenido algún impacto, aunque no llegó a inclinar el resultado final.

Periodismo e *influencers:* el espacio informativo mundial

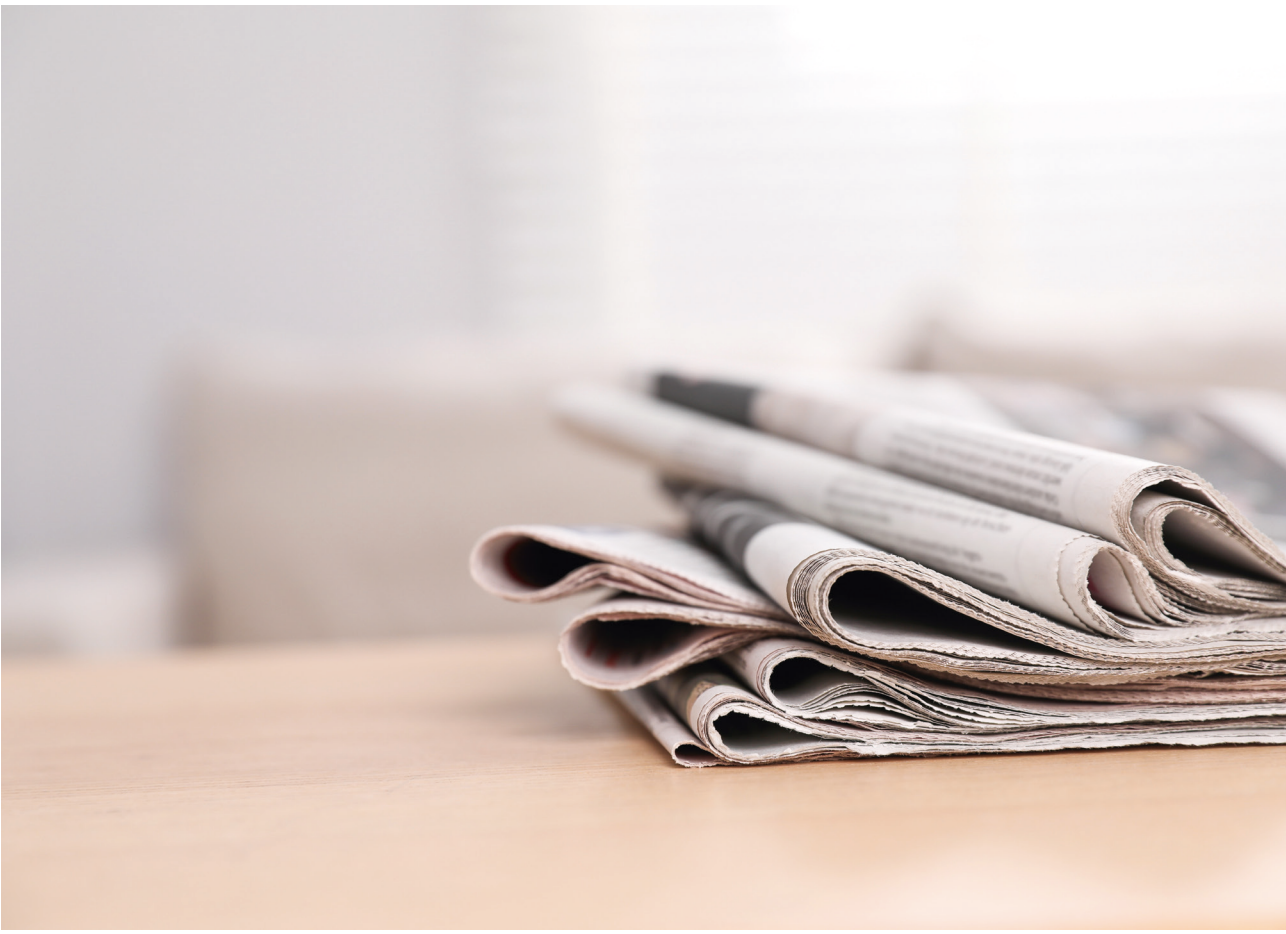
La IA generativa tiene el potencial de transformar completamente el panorama informativo global y afectar todas las formas de comunicación política, creación y presentación de contenidos, incluyendo el periodismo y el trabajo de los *influencers*.

IA y periodismo

La GenAI tiene un impacto significativo en la creación y presentación de contenidos en el ámbito del periodismo. Sin embargo, existe una aparente brecha de la IA: mientras los medios de comunicación tradicionales y de calidad se esfuerzan por encontrar respuestas, establecer límites y regular el uso ético de la IA en el periodismo, los medios de baja calidad, los tabloides, los medios de propaganda estatal y los estafadores ya están aprovechando la IA para sus propios fines. El medio ruso de propaganda extranjera RT, por ejemplo, ya utiliza personajes *deepfake*, es decir, avatares automatizados generados completamente por IA, a los que llaman presentadores digitales, para su programa en español. También se sabe desde hace tiempo que varios canales de noticias de China y otros países han estado utilizando la IA con estos fines.

Mientras que los medios tradicionales de calidad de todo el mundo suelen abstenerse de utilizar GenAI para crear noticias centrales y continúan haciéndolo por sí mismos, otros actores son más propensos a emplearla. Las agencias de noticias han identificado miles de sitios web que utilizan IA (principalmente ChatGPT) para gestionar sitios de noticias totalmente automatizados. Estos sitios web se dedican tanto a republicar y reescribir contenidos antiguos para generar ingresos publicitarios, como a difundir directamente desinformación política. Por otro lado, Channel 1, una nueva emisora de noticias establecida en Los Ángeles en 2024, es el primer medio que afirma ser un actor serio en el periodismo pero que gestiona íntegramente sus programas con GenAI, tanto en la creación de contenidos como en su presentación.

Otro aspecto importante de GenAI en el periodismo es cómo las platafor-



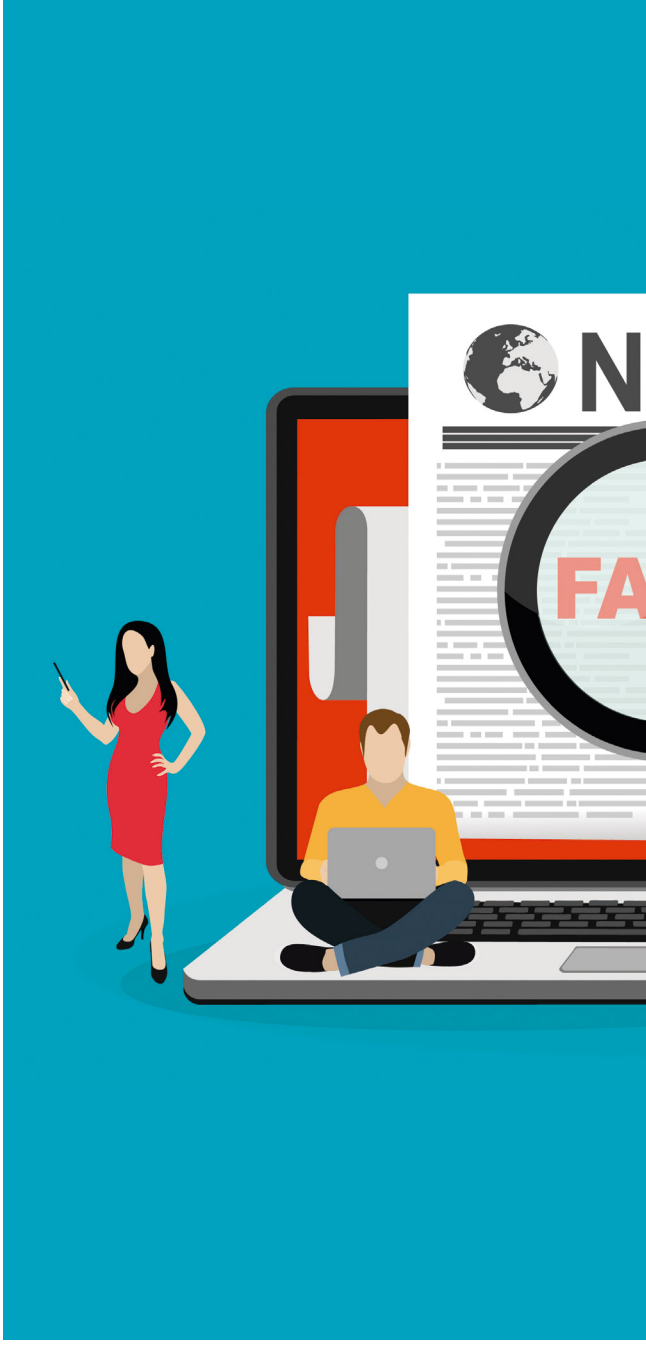
mas de redes sociales regulan, marcan y publican contenidos generados por IA. Aunque pronto será obligatorio que las plataformas marquen y especifiquen los contenidos generados por IA en América del Norte y la UE, no existen normas unificadas de este tipo en otras partes del mundo. La mayoría de las plataformas sociales afirman en sus normas comunitarias y condiciones de uso que los contenidos y perfiles generados por IA deben estar claramente marcados y registrados. Sin embargo, como en el pasado, el grado de cumplimiento de estas normas varía considerablemente.

IA e influencers

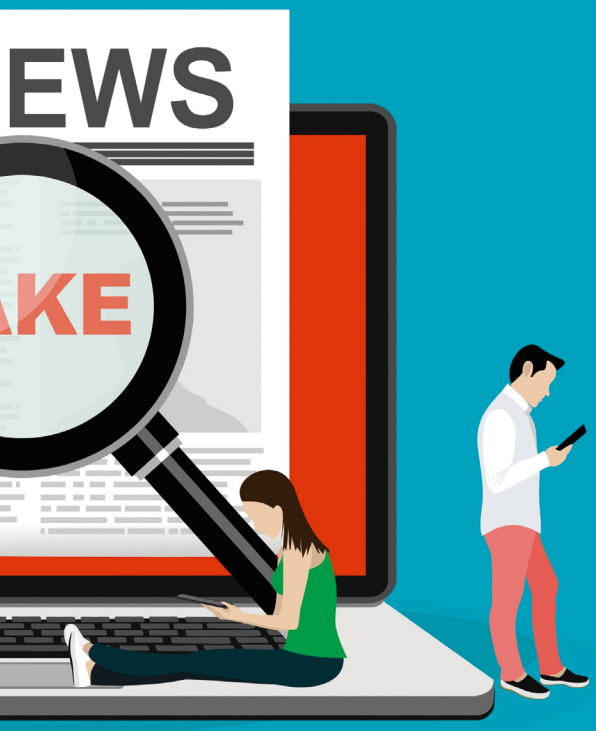
En el mundo de los *influencers* parece estar ocurriendo una evolución similar con la generación y presentación de contenidos, que son profundamente impactadas por GenAI. Los *influencers* virtuales, es decir, avatares creados y gestionados completamente por IA que se hacen pasar por *influencers* en plataformas de redes sociales, ya han atraído a millones de seguidores en países como China, Brasil, Estados Unidos e India. Esto también impacta la difusión de misinformación, desinformación, teorías conspirativas, etc. En todo el mundo,

los *influencers* están adquiriendo cada vez más relevancia, en tanto grupo objetivo y herramienta para los actores profesionales de la desinformación, así como creadores y difusores de desinformación.

Algunos *influencers* de TikTok, por ejemplo, han convertido en su modelo de negocio la creación de videos generados por IA sobre nuevas teorías conspirativas, y debaten en chats cerrados sobre cómo utilizar GenAI para incrementar sus ingresos. En otros casos, se ha demostrado que embajadas rusas en África pagan a *influencers* locales para difundir desinformación. Hasta ahora, no se ha comprobado que los *influencers* generados por IA estén involucrados en campañas políticas o en la difusión de desinformación, aunque el riesgo sigue siendo alto. Las manipulaciones de IA y los *deepfakes* también afectan a los profesionales de los medios y a los *influencers*: ambos grupos son frecuentemente víctimas de ataques de desprestigio mediante estas tecnologías. Los videos deepfake en los que periodistas promocionan, sin su consentimiento, estafas financieras o productos dudosos se han vuelto comunes en Estados Unidos y Europa. Los *influencers*, por su parte, a menudo enfrentan el riesgo de ser víctimas de *deepfakes* que dañan su reputación (y, en consecuencia, su modelo de negocio). El escenario más común en este caso es el uso de imágenes de *influencers* mujeres en *deepfakes* de contenido pornográfico. Estos ataques también pueden ser parte de campañas de desprestigio con fines políticos, como se vio con los *deepfakes* de Taylor Swift tras su involucramiento en elecciones estadounidenses de 2024.



» Rusia es considerada uno de los actores más activos en la manipulación e injerencia de información extranjera. «



» La inteligencia artificial está transformando de manera radical el panorama de la desinformación, la interferencia electoral y la manipulación de la información. «

Sobrecontaminación: ahogándonos en un mar de contenidos generados por IA

Otra dimensión de GenAI en el ámbito global de la información es la posibilidad real de su sobrecontaminación con contenidos generados por IA, robots automatizados y similares. Los escenarios pesimistas sugieren que, para 2026, el noventa por ciento de todos los contenidos en línea podrían ser generados por IA. Además, algunos estudios indican que el comportamiento automatizado en línea, como bots y programas, ya constituye la mayoría de las actividades en la red. Si la GenAI llega a dominar la mayor parte del contenido, la presentación y las actividades en línea, esto impactará gravemente en las noticias políticas y en la información en general y afectará a las sociedades. Por lo tanto, la sobrecontaminación podría convertirse en uno de los riesgos más serios de la GenAI a largo plazo.

Manipulación e injerencia de información extranjera rusa

Rusia es considerada uno de los actores más activos en la manipulación e injerencia de información extranjera (FIMI). La difusión coordinada y encubierta de información falsa, engañosa y manipulada es una herramienta clave para influir en sociedades, eventos y elecciones en estas actividades. La desinformación rusa está presente en casi todas las elecciones del mundo y utiliza una amplia gama de herramientas e instrumentos complejos. Las

embajadas y consulados rusos, medios de comunicación, empresas de relaciones públicas, periodistas *freelance, influencers*, portales web anónimos y proxis locales son los principales actores en la desinformación rusa. Sus tácticas abarcan desde la propaganda básica y el pago a *influencers* y periodistas hasta operaciones de desinformación más complejas que incluyen la falsificación de medios de comunicación respetados y la difusión encubierta de noticias falsas. Las narrativas y mensajes de la desinformación rusa suelen centrarse en temas clave (p. ej., antioccidentalismo, anti-Ucrania y anti-LGTBQ) que se adaptan en mensajes personalizados para audiencias locales en diferentes partes del mundo. En el Sur Global, estas narrativas suelen enfocarse en desacreditar a Occidente (p. ej., colonialismo, tensiones sociales, injusticia económica y social).

Usos de la IA

En sus operaciones de desinformación, los actores rusos de la FIMI han demostrado su disposición a explotar el potencial ilimitado de la GenAI para sus objetivos. El programa en español de la emisora rusa de propaganda extranjera RT ahora incluye dos presentadores digitales, es decir, avatares generados por IA. En Estados Unidos, varios sitios web de noticias falsas que publicaron artículos negativos completamente automatizados sobre las elecciones presidenciales, escritos con GenAI, tienen vínculos con Rusia. Durante una sofisticada campaña mundial de desinformación llamada Doppelganger, que se enfoca en sitios web falsificados de los medios tradicionales más conocidos, se descubrió que agentes rusos usaban ChatGPT para generar y traducir publicaciones y comentarios en redes sociales.

» Las estrategias basadas en IA pueden utilizarse para dirigir mensajes a votantes específicos, personalizar el contenido y aumentar la eficacia de las campañas políticas. «

En la guerra en curso contra Ucrania, los actores rusos han utilizado repetidamente videos *deepfake* (p. ej., uno falso del presidente Zelensky pidiendo la rendición y otro fabricado de los jefes de inteligencia ucranianos supuestamente admitiendo su participación en un atentado terrorista en Moscú) para desinformar tanto a la audiencia interna como externa.

Conclusiones

La inteligencia artificial está transformando de manera radical el panorama de la desinformación, la interferencia electoral y la manipulación de la información. Hoy en día, todas estas actividades malintencionadas han integrado la IA, y ya no hay elecciones sin algún nivel de desinformación generada por esta tecnología. La IA facilita la producción y difusión de este tipo de contenido a una velocidad sin precedentes, reduciendo costos y simplificando su creación. Esto ha permitido que las campañas de desinformación sean más accesibles, automatizadas, personalizables, persuasivas y de mayor alcance.

A pesar de estos avances, el temido *apocalipsis de la información* aún no se ha materializado. Ninguna elección ha sido

influida de forma decisiva por la desinformación generada por la IA, aunque ha habido casos notables en los que los *deepfakes* han jugado un papel significativo. Por ejemplo, en las recientes elecciones en Turquía y Eslovaquia, aunque los *deepfakes* llamaron la atención y suscitaron preocupación, al final no alteraron los resultados a favor de ningún candidato o partido.

Mientras tanto, la IA no es solo una herramienta para la desinformación, sino también una fuerza creciente en las campañas políticas. Las estrategias basadas en IA pueden utilizarse para dirigir mensajes a votantes específicos, personalizar el contenido y aumentar la eficacia de las campañas políticas. A medida que esta tendencia crece, también aumentan los riesgos asociados a las manipulaciones con IA, especialmente los *deepfakes*. Más allá de las elecciones, los *deepfakes* alimentan el ciberacoso, el fraude, las estafas y las violaciones de ciberseguridad, con los *influencers* siendo particularmente vulnerables a estos usos maliciosos de la tecnología de IA, como el ciberacoso mediante pornografía *deepfake.*

Bibliografía

Bontcheva, K. (ed.) (2024). Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities. https://www.veraai.eu/posts/white-paper-generative-ai-and-disinformation

Ferrara, E. (2024). GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models. https://DOI.org/10.48550/arXiv.2310.00737

Gehringer, F. A., Nehring, Ch., y Łabuz, M. (2024, 10 de mayo). The influence of Deep Fakes on Elections: Legitimate Concern or Mere Alarmism? *KAS Monitor 2024.* https://www.KAS.de/en/monitor/detail/-/content/the-influence-of-deep-fakes-on-elections

Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. *Synthese,* v. 201 103/2023. https://link.springer.com/article/10.1007/s11229-023-04097-3

Labuz, M., Nehring, Ch. (2024, 26 de abril). On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *Eur Polit Sci.* https://doi.org/10.1057/s41304-024-00482-9

Marchal, N., y Xu, R. (2024, 2 de agosto). Mapping the misuse of generative AI. *GoogleDeepmind.* https://deepmind.google/discover/blog/mapping-the-misuse-of-generative-ai/

Muñoz, M. (2024). *The AI Election Year: How to Counter the Impact of Artificial Intelligence.* DGAP Memo, v. 1. https://dgap.org/en/research/publications/ai-election-year

Schick, N. (2020). Deep Fakes and the *Infocalypse.* Ottawa.

Traducción inglés-español: Doris Filipovic.

Christoph Nehring

Investigador, analista y periodista. Profesor invitado y analista en el programa de medios de comunicación de la Fundación Konrad Adenauer, autor para *Tagesspiegel,* Deutsche Welle, *NZZ*, *Spiegel* y muchos otros. Apasionado de la IA y la desinformación. Lleva más de diez años investigando la desinformación, la manipulación y los servicios secretos.

LI: christopher-nehring-423b06257