Cuánto puede impactar la IA en la desinformación: el periodismo en la era de los 'deepfakes'

Al principio pensábamos que bastaba con mirar los dedos de las manos para detectar a la máquina. Pero luego **los *deepfakes* fueron invadiendo nuestros chats y redes sociales** y nos dimos cuenta de que en poco tiempo esto ya no bastaría. ¿Cómo creer que algo es cierto cuando no puedes fiarte de tus propios ojos? En un entorno informativo cada vez más inundado de contenido sintético, periodistas y comunicadores, nos preguntamos cada vez más cómo enfrentarnos a **un mundo de contenido creado con inteligencia artificial** (IA).

MARILÍN GONZALO

Aunque la imagen del papa con un abrigo inflable podía parecernos una diversión inofensiva en Twitter (ahora, X), hoy prácticamente no hay campaña electoral o crisis política en la que no aparezcan imágenes de políticos manipuladas, con la consiguiente confusión desinformada. En los primeros días de la campaña para las elecciones en Cataluña, Ciudadanos colgó carteles con una imagen[1] en la que Pedro Sánchez y Carles Puigdemont se dan la mano, debajo de un texto en mayúsculas que decía "Detenlos". La foto nunca había sucedido, y abajo, varias decenas de puntos de la fuente más pequeña, el texto expiatorio: "Imagen generada por inteligencia artificial". Un vídeo de la misma campa-

1 JorgeCarrion21(26 de abril de 2024). Ciudadanos usa una imagen *fake*, generada por inteligencia artificial, en su campaña electoral [*Post*]. https://x.com/jorgecarrion21/status/1783751982269800876?s=12&t=dsiEnuFEnWguQy-w3F-HJw

Marilín Gonzalo es periodista especializada en medios digitales, escribe sobre tecnología en su intersección con medios y derechos humanos. Trabaja en *Newtral*, donde lleva Políticas Públicas y escribe una columna sobre tecnología

ña[2] creado de forma sintética mostraba una pancarta gigante que se despliega en una plaza de Barcelona reproduciendo ese cartel. Esta vez no había ninguna leyenda que advirtiera del carácter artificial de la foto, como tampoco lo hubo en varios *posts* que reprodujeron la foto y el vídeo falsos.

Los avances de la IA nos han sorprendido en los últimos años por su rapidez y por su impacto en casi todos los ámbitos. La IA generativa apunta especialmente a los espacios de creación de contenidos y, aunque todavía muchos hablan de los riesgos en futuro, la inteligencia artificial ya está provocando daños a niveles más diarios y cercanos.

El primer uso masivo que tuvieron los *deepfakes* en vídeo fue crear contenido machista, y así lo advirtió un estudio de Deeptrace que aseguraba que el 96% de los vídeos *deepfakes* son pornográficos y no consentidos. "La pornografía con IA daña exclusivamente a mujeres", decía aquel informe de 2019. Sin embargo, no pensamos en esto cuando hablamos del daño que causa la desinformación, sino más bien en las noticias falsas o narrativas negacionistas.

Además de generar violencia contra la mitad de la población, la IA generativa ha llenado los entornos digitales de vídeos acosadores y hechos manipulados. No solo eso, también de reseñas falsas en grandes tiendas *online*, bots que comentan o ponen *likes* en conversaciones llenas de mentiras, páginas web creadas con apariencia de medios de comunicación con el fin de conseguir clics en titulares sensacionalistas, que la IA también produce, integrada ya en programas de edición y SEO que utilizan medios de comunicación.

Además de generar violencia contra las mujeres, la IA generativa ha llenado los entornos digitales de vídeos acosadores y hechos manipulados

La fascinación producida por ChatGPT, MidJourney y desarrollos derivados de grandes modelos de lenguaje similares nos hace olvidar que son proclives a alucinaciones y dan datos que no pasan una verificación. Y los humanos no las hacemos. A medida que aumenta la calidad de la IA, tenemos menos incentivos para esforzarnos y permanecer atentos, lo que permite a la IA sustituir, en lugar de aumentar, su rendimiento. Ello lo descubrió Fabrizio Dell'Acqua, profesor en el Laboratorio para la Ciencia de la Inno-

2 VotaCiudadanos (26 de abril de 2024). Vídeo de la campaña de Ciudadanos [*Post*]. https://x.com/VotaCiudadanos/status/1783847316341789130

vación, de la Harvard Business School, que llamó a este fenómeno "quedarse dormido al volante" de la IA[3].

Otro estudio, este de Ethan Mollick, profesor de Wharton especializado en innovación y empresas, observó lo mismo: "Nuestra investigación muestra que las personas no hacen comprobaciones sobre el trabajo de las IA, una vez que estos sistemas pasan cierta línea de calidad, y el entrenamiento tampoco ayuda en este sentido"[4].

Cada vez más textos surgidos de chatbots pueblan discursos, plataformas de ventas, librerías de *ebooks*, estudios científicos y textos periodísticos, legales, de *marketing* o de divulgación. La desinformación producida por estas herramientas llega también al audio: en el contexto de las elecciones primarias del partido demócrata en EE. UU., un consultor político envió a entre 5.000 y 25.000 ciudadanos estadounidenses una llamada con la voz de Joe Biden que sugería falsamente que votar en estas elecciones impediría a los votantes hacerlo en noviembre.

La preocupación por los 'deepfakes' viene de lejos

Cuando los primeros *deepfakes* comenzaron a conocerse a través de un foro en Reddit a finales de 2017, Danielle Citron, profesora de derecho de la Universidad de Maryland, pensó que, además de atacar la privacidad de las mujeres, si se diseminaba esta tecnología, podía convertirse en un arma mucho más peligrosa de desinformación para debilitar sociedades democráticas. Junto con su compañero Robert Chesney, publicó un estudio en el que ya preveían que cualquiera con acceso a estas herramientas, desde propagandistas consentidos por el Estado hasta troles, podría sesgar la información, manipular las creencias y, al hacerlo, enfrentar a comunidades en línea ideológicamente opuestas hacia sus propias realidades subjetivas. "El mercado de las ideas ya sufre la decadencia de la verdad, puesto que nuestro entorno de información en red interactúa de forma tóxica con nuestros prejuicios cognitivos", según el informe. "Los *deepfakes* exacerbarán este problema de forma sig-

3 Dell'Acqua, F. (2022). "Quedarse dormido al volante: colaboración entre humanos y IA en un experimento de campo con reclutadores de recursos humanos". Laboratorio de Ciencias de la Innovación, Harvard Business School. https://static1.squarespace.com/static/604b23e38c22a96e9c78879e/t/62d5d9448d061f7327e8a7e7/1658181956291/Falling%2BAsleep%2Bat%2Bthe%2BWheel%2B-%2BFabrizio%2BDellAcqua.pdf

4 Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... y Lakhani, K. R. (2023). "Navegando por la irregular frontera tecnológica: evidencia experimental de campo de los efectos de la IA en la productividad y la calidad de los trabajadores del conocimiento". Gerente de Tecnología y Operaciones de la Escuela de Negocios de Harvard. Unidad de Tecnología y Gestión de Operaciones. Harvard Business School, n.º 24-013. http://dx.doi.org/10.2139/ssrn.4573321

nificativa", concluyó.

Ese mismo año, vimos a un Barack Obama diciendo en YouTube que necesitábamos estar alerta en internet. "Cómo avancemos en la era de la información marcará la diferencia entre que sobrevivamos o nos transformemos en una jodida distopía. Gracias, manteneos despiertos, putas". A pesar de las palabras malsonantes y ciertos errores visuales que hoy reconocemos en la sincronía de los labios de la imagen de este Obama, la mayoría se lo creyó y el *deepfake* se hizo viral. Su creador, Jordan Peele, un realizador audiovisual estadounidense, lo había hecho junto con el director de *Buzzfeed*, Jonah Peretti, para llamar la atención al público sobre la amenaza de la desinformación digital, en un entorno de nuevas tecnologías y la erosión de una realidad compartida.

Algo parecido es lo que sucedió más recientemente con las falsas imágenes de Donald Trump saliendo de los tribunales detenido, las cuales probablemente todos tenemos aún en la retina, creadas por Eliot Higgins, fundador de Bellingcat, una organización de periodistas que hacen investigación con *fact-checking* e inteligencia de fuentes abiertas. Aunque fueron creadas con Midjourney y claramente marcadas como falsas, se hicieron virales rápidamente. La intención del investigador era también concienciar sobre el poder de estas herramientas, aún en su infancia.

Un año después del episodio, pregunto a Higgins si piensa que ha cambiado algo: "Creo que la gente es más consciente de lo que pueden hacer las imágenes generadas por IA y de lo convincentes que resultan, pero sigo pensando que, a medida que los sistemas de inteligencia artificial se desarrollen con rapidez, habrá una brecha de conocimiento entre lo que el público sabe de lo que es capaz la IA y lo que puede hacer en realidad. Esta brecha es la que abre la posibilidad de que los actores más nefastos cometan maldades, por lo que contar con un público informado es una parte importante de la capacidad de resistencia frente a la manipulación".

Para desactivar la desinformación generada por IA no hay nada más efectivo que el pensamiento crítico del público

Para desactivar este tipo de desinformación no hay nada más efectivo que el pensamiento crítico del público, y tenemos un buen ejemplo con lo que probablemente fue el primer *deepfake* en una guerra. Sucedió a las pocas semanas del inicio de la invasión de Ucrania, pero fue tan rápidamente desmontado que a muchos no les dio tiempo a creérselo. Un vídeo de un falso Volodímir Zelenski se dirigía a su nación para anunciar la rendición, algo que había sido previsto

por el hábil Centro de Comunicación Estratégica del Gobierno ucraniano, que ya había distribuido información a la ciudadanía sobre estas tecnologías, la posible aparición de *deepfakes* y, en concreto, de uno en el que Zelenski anunciara su capitulación en pleno conflicto. El mismo Zelenski, que tiene comunicación directa con la ciudadanía a través de su canal de Telegram, salió a desmentirlo minutos después de su aparición.

Solo unos meses después, una serie de videollamadas mantenidas por un falso alcalde de Kiev con cinco alcaldes europeos, entre ellos el de Madrid, José Luis Martínez-Almeida, levantaban alarmas. El alcalde de Viena, Michael Ludwig, estaba tan convencido de haber hablado con Vitali Klitschko, el original kievita, que al terminar tuiteó y emitió una nota de prensa dando cuenta de lo que se había hablado, con imágenes de la videoconferencia.

Los primeros en contar el caso y decir que "parecían estar tratando con un *deepfake*" fueron los alemanes, cuya alcaldesa habló durante 15 minutos con el falso Klitschko, hasta que empezó a sospechar por la índole de las preguntas, enfocadas hacia los refugiados ucranianos en Berlín. Las videollamadas eran similares a las que mantuvieron los regidores de Varsovia, Viena, Budapest y Madrid. Un dúo de cómicos rusos se atribuyó la autoría de estas llamadas, atribuyéndoles un fin de entretenimiento. El tema no causó mucha gracia: la situación de guerra en Europa no estaba para chistes y los cómicos nunca presentaron pruebas. Quedaron dudas sobre las verdaderas intenciones de estos engaños y si el Gobierno ruso estaba implicado en episodios que buscaban dejar en ridículo a los alcaldes occidentales.

Desinformación con IA en elecciones

Si las decisiones de una persona se basan en la información que tiene, si la democracia se sostiene en las decisiones informadas de los votantes, ¿cómo afectará una evolución de estas tecnologías en procesos electorales políticos, capaces de modificar el destino de poblaciones enteras? Las campañas de noticias falsas, para tener más efecto, tienden a centrarse en periodos cortos y críticos, como las crisis y las elecciones, apuntando a grupos determinados que puedan virar las encuestas.

¿Puede la desinformación generada por IA impactar en los resultados electorales? La pregunta cobra relevancia en un año en el que unos 2.000 millones de personas están llamados a votar en unas 70 elecciones alrededor del mundo. La polarización y el avance de la ultraderecha parecen configurar la tormenta perfecta y es complicado determinar el impacto de unas tecnologías cuando hay tantas variables en juego.

Las últimas elecciones presidenciales en Argentina se convirtieron en un campo de pruebas para la IA, con los partidos de Javier Milei y Sergio Massa apli-

cándola para crear narrativas propias. Los carteles de campaña presentaban a los candidatos como próceres, o a los contrarios como monstruos, caricaturas o personajes malvados de la cultura popular. No obstante, la mayor cantidad de estas imágenes no provino de los equipos de los candidatos, sino de cuentas no oficiales de militantes y partidarios, en algunos casos con más seguidores. El uso de la IA generativa en esta campaña ha sido un fenómeno pequeño, pero va a ser mayor, según Contextual, una organización que monitorea desinformación y discursos antidemocráticos. "Hay herramientas económicas al alcance de cualquiera. Hay inversión por parte de los equipos de campaña, pero también hay movilización activista y no rentada", expusieron.

Estudios señalan que el fin de las campañas de desinformación no es que te creas algo, sino que no creas en nada

Aunque un puñado de imágenes causaron desinformación, el contenido generado por IA estaba etiquetado como tal o era una falsificación tan evidente que es poco probable que haya engañado a los votantes, según Jack Nicas y Lucía Cholakian Herrera, en un reportaje del *New York Times* desde Buenos Aires.

Según Javier Pallero, director de Contextual, iniciativa del Instituto de Desarrollo Digital de América Latina y el Caribe para añadir análisis a contextos políticos, lo que más circula en Argentina son *cheapfakes*, contenidos creados con fallas muy visibles, que, sin embargo, pueden llegar a tener un impacto. "A veces no es necesario un gran despliegue para que la gente lo crea. Este es un tema que atraviesa todo el problema de la desinformación: no hace falta que sea convincente la noticia falsa, sino que tiene que abonar una teoría que ya se venía trabajando en ese público que está ansioso de ver algo horrible sobre el candidato que rechaza", explicó a *Perfil*.

Crear desconfianza y ciudadanos desinformados

En México también 2024 es año electoral. La desinformación tradicional alrededor del voto es mucho más frecuente que la generada mediante IA, dijo Arturo Daen, editor de la sección de verificación de *Animal Político*. El realismo de las imágenes, en este sentido, no es algo que necesariamente la gente busque o que sea condición para que algo sea viral, explicó Sacha Altay, investigadora posdoctoral cuyo campo actual es la desinformación, la confianza y los medios sociales en el Laboratorio de Democracia Digital de la Universidad de Zúrich.

Altay, junto con investigadores como Felix Simon y Hugo Mercier, coincidieron en que la preocupación por que la IA

agrave el problema de la desinformación es exagerada. En un artículo para Harvard Kennedy School, destacaron que la IA puede aumentar la cantidad de desinformación, pero esto no significa necesariamente que la gente la consuma más, debido a los límites de la demanda de desinformación, que ya es abundante. Aunque aumente la calidad de la desinformación, es posible que tampoco tenga un gran impacto en el público, puesto que ya existen herramientas ajenas a la IA -pensemos en Photoshop o editores de vídeo- que pueden hacer que las imágenes falsas parezcan realistas. Además, la persuasión es compleja,por lo que incluso la desinformación personalizada de alta calidad tendría probablemente un impacto limitado.

Cuando analizamos desinformaciones, siempre quedan dudas sobre las verdaderas intenciones detrás de estos engaños. Muchos estudios señalaron que el fin último de las campañas de desinformación no es que te creas algo, sino que no creas en nada. En este sentido, el uso de la IA generativa parece ayudarles, como está sucediendo con la guerra de Gaza.

Los observadores de la desinformación han acertado en su predicción de que la tecnología se impondría en la guerra, pero no exactamente por la razón que pensaban, según varios medios como el *New York Times* y *Wired*. "No estamos viendo un uso masivo de imágenes generadas por IA en el conflicto entre Israel y Hamás y la desinformación relacionada con ello", aclaró Tomasso Canetta, del Observatorio de Medios Digitales Europeos. Los investigadores han encontrado relativamente poco contenido falso generado por IA, y aún menos, convincente. No obstante, la mera posibilidad de que puedan estar circulando contenidos generados por IA está llevando a la gente a descartar imágenes, vídeo y audio auténticos. Recuerda: no que se crean algo, sino que no crean nada. Esto también es desinformación.

Menos 'apps' y más periodismo y alfabetización

¿Cuál es la mejor herramienta para detectar *deepfakes* o contenido generado por IA? En estos tiempos de solucionismo tecnológico, esta es una de las preguntas que más recibimos los verificadores. Como si el tremendo problema de la desinformación se solucionara con una *app*.

Aun así, hay empresas tecnológicas que trabajan en encontrar esa solución, si bien lo cierto es que, a pesar de titulares llamativos y esfuerzos de *marketing*, hasta la fecha ninguna herramienta creada para detectar contenido sintético ha conseguido resultados que la hagan fiable *per se*. El mismo OpenAI abandonó el desarrollo de su detector de contenido generado por IA tras constatar que no llegaban a una precisión en los resultados mínimamente satisfactoria. La evolución de la IA generativa siempre estará un paso por delante, pues es la misma

industria: quienes quieran intentar optimizar su modelo mirarán lo que los detectores -humanos o máquinas- miren: los dedos, las sombras o los dientes, y eso es lo que mejorarán con la próxima actualización.

Menos contar dedos en imágenes realistas y más pensar qué emociones nos quiere causar una imagen, qué intenta provocar si ese relato se viraliza

Lo primero que hacen los investigadores dedicados a detectar la fiabilidad de imágenes es periodismo, por decirlo brevemente. Sam Gregory es un experto en desinformación, y dirige Witness, una organización por los derechos humanos enfocada en el vídeo, que ha puesto un grupo de expertos en *deepfakes* a disposición de verificadores en todo el mundo. Me explica que, cuando sospechan que están ante una imagen creada por IA, el primer paso es "parar, investigar la fuente, buscar coberturas alternativas y rastrear al original antes de dejar que las emociones o primeras impresiones nos dirijan".

Cuando le pregunto por detalles concretos para buscar, responde que el estado actual de la IA crea problemas en la representación de las manos, como distorsión, o en las sombras y las perspectivas en la imagen, pero advierte que no debemos confiar en esto. "Los sistemas avanzan rápidamente, mejoran y apuntan a una representación más realista. Sabemos por experiencia previa que estas pistas pueden desaparecer rápidamente; por ejemplo, se solía pensar que los *deepfakes* no pestañeaban y ahora lo hacen".

Los expertos creen que la rápida evolución de la IA hace creíbles sus riesgos, aunque alertan de que no debemos olvidar la capacidad de las personas para aprender a identificar este material, sobre todo cuando tienen acceso a estas herramientas. La alfabetización digital, y más específicamente sobre IA, se ha mostrado en todos los estudios a lo largo de los últimos años como un remedio eficaz contra la desinformación.

Gregory dice que los periodistas y activistas de distintos países con los que trabaja coinciden en que quienes más se benefician de socavar la verdad y nuestra confianza en las imágenes son las personas en el poder que quieren debilitar los relatos críticos y el "periodismo ciudadano". El investigador apunta al fenómeno conocido como "dividendo del mentiroso", por el que los desinformadores afirman que imágenes y vídeos que sí son reales podrían ser *deepfakes* y, por lo tanto, no ser dignos de confianza.

Menos contar dedos en imágenes realistas y más pensar qué emociones nos quiere causar una imagen, qué intenta provocar si ese relato se viraliza. "Si no

tenemos cuidado, la exageración y el pánico pueden causar daños reales y socavar las voces más importantes de nuestras sociedades", avisa Gregory. ■

Bibliografía

- Adami, M. (15 de marzo de 2024). "Cómo la desinformación generada por IA podría afectar las elecciones de este año y cómo los periodistas deberían informar al respecto". Instituto Reuters. https://reutersinstitute.politics.ox.ac.uk/news/how-ai-generated-disinformation-might-impact-years-elections-and-how-journalists-should-report
- Bedingfield, W. (30 de octubre de 2023). "La IA generativa está desempeñando un papel sorprendente en la desinformación entre Israel y Hamás". *Wired*. https://www.wired.com/story/israel-hamas-war-generative-artificial-intelligence-disinformation/
- Eisele, I. (11 de octubre de 2023). "Verificación de hechos: falsificaciones de IA en la guerra de Israel contra Hamás". *DW*. https://www.dw.com/en/fact-check-ai-fakes-in-israels-war-against-hamas/a-67367744
- Entrevistas a Sam Gregory, director de Witness, y Eliot Higgins, fundador de Bellingcat. Gonzalo, M. (10 de marzo de 2021). *"Deepfakes,* mentiras y vídeo". *Newtral*. https://www.newtral.es/que-son-deepfakes-inteligencia-artificial/20210310/
- Gonzalo, M. (18 de marzo de 2022). "Cómo se desmontó el *deepfake* de Zelenski, el primero de la guerra contra Ucrania". *Newtral.* https://www.newtral.es/deepfake-zelenski-primer-guerra-contra-ucrania/20220318/
- Gonzalo, M. (7 de julio de 2022). "Cómo se hace un *deepfake* del falso alcalde de Kiev para engañar a cinco regidores europeos". *Newtral*. https://www.newtral.es/falso-alcalde-kiev-deepfake-almeida-klitschko/20220707/
- Hsu, T. y Thompson, S. (28 de octubre de 2023). "La IA enturbia la guerra entre Israel y Hamás de manera inesperada". *The New York Times.* https://www.nytimes.com/2023/10/28/business/media/ai-muddies-israel-hamas-war-in-unexpected-way.html
- Leclercq, G. (13 de noviembre de 2023). "Inteligencia artificial en la campaña: el falso vídeo de Sergio Massa tomando cocaína disparó el debate". *Perfil*. https://www.perfil.com/noticias/politica/inteligencia-artificial-en-la-campana-el-falso-video-de-sergio-massa-tomando-cocaina-disparo-el-debate.phtml
- Nicas, J. y Cholakian Herrera, L. (15 de noviembre de 2023). "Las campañas electorales de Argentina recurren a la IA". *The New York Times"*. https://www.nytimes.com/es/2023/11/15/espanol/elecciones-argentina-imagenes-inteligencia-artificial.html
- Simon, F. M., Altay, S., y Mercier, H. (2023). "¿Se ha recargado la información errónea? Los temores sobre el impacto de la IA generativa en la desinformación son exagerados". *Misinformation Review*. Harvard Kennedy School (HKS). https://doi.org/10.37016/mr-2020-127