Cuadernos del Audiovisual
del Consejo Audiovisual de Andalucía

Deepfakes, desinformación, discursos de odio y democracia en la era de la Inteligencia Artificial

Deepfakes, disinformation, hate speech and democracy in the age of Artificial Intelligence

Aníbal M. Astobiza
Universidad del País Vasco / Euskal Herriko Unibertsitatea
Globernance-Instituto de Gobernanza Democrática
anibal.monasterio@ehu.eus https://orcid.org/0000-0003-1399-5388

Recibido: 22/03/2024 | Aceptado: 03/06/2024

Resumen: La inteligencia artificial (IA) está transformando el mundo y esto implica que también presenta nuevos desafíos. Una de las mayores preocupaciones es el auge de los *deepfakes,* vídeos o audios manipulados para que parezcan reales. Esta tecnología tiene el potencial de ser utilizada para difundir desinformación, propaganda y discurso de odio, lo que representa una grave amenaza para la democracia. Los *deepfakes* son cada vez más sofisticados y difíciles de detectar. Esto significa que pueden ser utilizados para manipular a la opinión pública de forma muy efectiva. Por ejemplo, se pueden crear *deepfakes* de políticos diciendo cosas que nunca han dicho o de celebridades haciendo cosas que nunca han hecho. La difusión de *deepfakes* puede tener un impacto devastador en la sociedad. Puede erosionar la confianza en las instituciones, aumentar la polarización política y alimentar la violencia. En un mundo donde la gente no sabe qué creer, la democracia se vuelve vulnerable. Pero, ¿es verdaderamente la IA una amenaza para la democracia? No lo creo. Este escenario no debe conducirnos a considerar la IA como una amenaza inexorable a la democracia. La clave reside en asumir la responsabilidad colectiva de desarrollar mecanismos efectivos de detección, promover la alfabetización digital y el pensamiento crítico entre la ciudadanía y salvaguardar el compromiso con la veracidad y el debate constructivo en el ámbito digital. En lugar de ceder al pánico ante la posibilidad de manipulación, debemos enfocarnos en empoderar a los individuos para que puedan navegar con discernimiento en este complejo ecosistema informativo. Este artículo propone, por tanto, una perspectiva esperanzadora: en la batalla contra el uso malintencionado de la IA para fines de desinformación y odio, el verdadero desafío podría no ser la tecnología en sí, sino nuestro propio temor. Superar este miedo mediante la educación, el desarrollo de habilidades críticas y el fomento de una cultura de verificación y participación activa, puede no solo mitigar los riesgos asociados a los *deepfakes*, la desinformación o el discurso de odio, sino también reforzar los pilares de nuestra democracia en la era digital.

Palabras clave: *deepfakes*; inteligencia artificial; desinformación; propaganda; discurso de odio; democracia.

Abstract: Artificial intelligence (AI) is transforming the world, and this implies that it also presents new challenges. One of the biggest concerns is the rise of deepfakes, vídeos s or audios manipulated to appear real. This technology has the potential to be used for spreading misinformation, propaganda, and hate speech, posing a serious threat to democracy. Deepfakes are becoming increasingly sophisticated and difficult to detect. This means they can be used to manipulate public opinion very effectively. For

Cómo citar:

Astobiza, A.M., (2024). Deepfakes, desinformación, discursos de odio y democracia en la era de la IA *Cuadernos del Audiovisual del Consejo Audiovisual de Andalucía*, (12), 177-190. https://doi.org/10.62269/cavcaa.20

Junta de Andalucía ca:A Consejo Audiovisual de Andalucía

example, deepfakes of politicians saying things they have never said, or celebrities doing things they have never done, can be created. The spread of deepfakes can have a devastating impact on society. It can erode trust in institutions, increase political polarization, and fuel violence. In a world where people do not know what to believe, democracy becomes vulnerable. But is AI truly a threat to democracy? I do not think so. This scenario should not lead us to view AI as an inexorable threat to democracy. The key lies in taking collective responsibility to develop effective detection mechanisms, promote digital literacy and critical thinking among citizens, and safeguard a commitment to truthfulness and constructive debate in the digital realm. Rather than succumbing to panic at the possibility of manipulation, we should focus on empowering individuals to navigate this complex information landscape with discernment. This article proposes, therefore, a perspective of hope: in the battle against the malicious use of AI for purposes of disinformation and hatred, the real challenge might not be the technology itself, but our own fear. Overcoming this fear through education, the development of critical skills, and the promotion of a culture of verification and active participation can not only mitigate the risks associated with deepfakes, misinformation, or hate speech, but also strengthen the pillars of our democracy in the digital age.

Keywords: deepfakes; artificial intelligence; misinformation; propaganda; hate speech; democracy.

1. Introducción

A principios de 2024, la cantante norteamericana Taylor Swift fue víctima de una serie de *deepfakes,* imágenes manipuladas que la hacían parecer desnuda o en situaciones comprometedoras (Saner, 2024). Las imágenes se difundieron principalmente en la plataforma X. La reacción de la plataforma fue tan lenta que una de las imágenes acumuló 47 millones de visitas antes de ser eliminada. Fueron principalmente los fans de Swift quienes se movilizaron y denunciaron masivamente las imágenes. El caso generó indignación pública, e incluso la Casa Blanca lo calificó de «alarmante». Finalmente, X eliminó las imágenes y bloqueó las búsquedas del nombre de la estrella del pop el domingo por la noche.

La mayoría de la gente descubrió los *deepfakes* por primera vez en 2017. Todo empezó cuando un usuario anónimo de Reddit publicó unos vídeos en los que aparecían celebridades, como Scarlett Johansson, en situaciones sexuales comprometidas. Pero no eran imágenes reales, sino el rostro de la famosa superpuesto al cuerpo de una actriz porno, creado con tecnología *deepfake* para que pareciera algo real. Al principio, los más vulnerables a este tipo de engaño eran los famosos y las figuras públicas, ya que los algoritmos necesitaban una gran cantidad de material vídeos gráfico para generar un *deepfake*, algo que abundaba en el caso de «celebrities» y políticos.

Me gustaría ofrecer una definición de *deepfake* más formal. Un *deepfake* es una creación o alteración de contenido audiovisual, en el cual las técnicas de aprendizaje profundo, particularmente las redes neuronales generativas antagónicas (GANs), se emplean para reemplazar la apariencia de una persona en un vídeo existente por la de otra persona (Chesney y Citron, 2019). Estas manipulaciones son a menudo tan precisas que pueden engañar al ojo humano y a las máquinas diseñadas para detectar manipulaciones, generando preocupaciones significativas sobre su uso para la desinformación y la manipulación mediática (Paris y Donovan, 2019).

La llamada inteligencia artificial (IA) generativa tiene sus raíces en los años cincuenta con la investigación en redes neuronales. Sin embargo, no fue hasta finales del siglo XX y

ISSN: 2952-6094 | e-ISSN: 2952-6116 | Núm. 12 | Julio de 2024
https://doi.org/10.62269/cavcaa.20 | Consejo Audiovisual de Andalucía

principios del XXI que la tecnología comenzó a avanzar a pasos agigantados gracias a la mayor potencia computacional y la disponibilidad de grandes conjuntos de datos. A partir de la década de 2010, los modelos generativos como GANs (Generative Adversarial Networks) y VAEs (Variational Autoencoders) experimentaron un rápido desarrollo en el ámbito académico, con aplicaciones en áreas como la generación de imágenes, música y texto.

La IA generativa comenzó a llegar al público general a partir de 2017, con la aparición de aplicaciones como Prisma, que permitía convertir fotos en pinturas al estilo de Van Gogh. La tecnología también se utilizó para crear *deepfakes*, vídeos manipulados que podían ser usados para desinformar o suplantar la identidad de personas. Pero ha sido en los últimos años cuando empresas como Google, Microsoft y OpenAI han lanzado productos y servicios que integran la IA generativa. Esta IA generativa permite realizar múltiples tareas de manera automatizada y como consecuencia incrementa la productividad. Pero la IA generativa tiene un lado oscuro. Es capaz de crear *deepfakes*.

La cuestión es: ¿podemos imaginarnos un escenario donde el espacio público de debate, los medios de información de masas, como por ejemplo, plataformas de redes sociales, estén inundados de contenido falso que incite la desinformación, difamación o suplantación de identidades? ¿Podemos imaginarnos Internet lleno de bulos, discursos y de odio y ciberacoso?

Es indudable que la IA generativa, y la IA en general, tienen el potencial de transformar nuestras sociedades de maneras inimaginables. Sin embargo, surge una pregunta crucial: ¿Podría esta tecnología suponer una amenaza para la democracia? Si bien la IA generativa presenta riesgos para la democracia, también ofrece oportunidades para fortalecerla. La clave está en abordar los riesgos de manera proactiva y aprovechar las oportunidades de forma responsable.

En este artículo, pretendo abordar la preocupación que generan las tecnologías de IA generativa, responsables de los *deepfakes* y la desinformación. Abordaré este tema desde un enfoque que combina el análisis conceptual y teórico con una sólida base empírica (Mercier y Sperber 2017; Mercier y Sperber, 2011; McKay y Dennett, 2009; Bronstein, Pennycook, Bear, Rand y Cannon, 2019). Esta metodología integral permitirá una comprensión profunda y matizada del asunto. Porque el análisis conceptual y teórico y los estudios empíricos son dos pilares fundamentales de la investigación científica**.** Ambos se complementan y fortalecen mutuamente, permitiendo obtener una comprensión más completa y rigurosa de cualquier tema.

Si bien es cierto que estas tecnologías (*e.g.* tecnologías como la IA y conexas que pueden crear *deepfakes*) pueden suponer un riesgo, es importante analizar la situación con detenimiento y evitar caer en alarmismos o pánicos morales. La desinformación no es un problema nuevo. A lo largo de la historia, se han utilizado diversos métodos para difundir información falsa, desde la propaganda política hasta el simple rumor. La digitalización ha exacerbado este desafío, simplificando la generación y propagación de información falsa y discursos de odio en una escala masiva. No obstante, es improbable que esto constituya una amenaza significativa para nuestras democracias.

Para fundamentar este enfoque, me respaldaré en cuatro argumentos proporcionados por el filósofo y actualmente profesor en la Universidad de Sussex, Dan Williams, especialista en ciencias cognitivas, filosofía de la mente y epistemología social, quien

recientemente ha centrado sus investigaciones en la desinformación en contextos digitales. Williams (2024) dice:

1) La desinformación en línea no constituye la raíz de los problemas políticos modernos.
2) La persuasión política es extremadamente difícil.
3) El entorno mediático es altamente competitivo y movido por la demanda.
4) El *establishment* tendrá acceso a formas de IA más poderosas que las fuentes contrarias al *establishment*.

Sobre la primera razón, Williams afirma que, desde 2016, prevalece la narrativa de que la desinformación en línea es una causa fundamental de diversos problemas políticos en democracias occidentales, como el auge de movimientos populistas de derecha, la disminución de la confianza en instituciones y el rechazo a consejos de salud pública. Esta narrativa ha variado en sus enfoques y en los «villanos» identificados (campañas de desinformación rusas, noticias falsas, Cambridge Analytica, «El Algoritmo», teorías de conspiración en línea, burbujas de filtro, etc.). Sin embargo, el argumento central de que la desinformación en línea impulsa muchos de los problemas políticos se afirma continuamente por expertos, científicos sociales, políticos, periodistas y documentales.

Las pruebas indican que la desinformación en línea no es más prevalente que en el pasado, y que, en algunos aspectos, los ciudadanos democráticos están mejor informados. Además, la mayoría de las personas siguen principalmente medios de comunicación convencionales, y aquellos que consumen contenido extremista o conspirativo en línea tienden a tener características y mentalidades específicas que buscan reafirmar sus identidades y visiones del mundo preexistentes. En consecuencia, la exposición a la desinformación en línea juega un papel menor en la formación de desconfianza y desdén hacia las instituciones establecidas, siendo estos sentimientos más bien el resultado de causas más complejas y profundas.

Sobre la segunda razón, Williams, apoyado en varios programas de investigación recientes en ciencias cognitivas, es mas contundente. Es extremadamente difícil cambiar las opiniones de las personas, y aún más cambiar sus mentes de maneras que conlleven a cambios en el comportamiento. Intuitivamente, reconocemos esta dificultad cuando debatimos en situaciones cotidianas, pero solemos olvidarla al considerar la persuasión en abstracto, cayendo en la creencia de que las personas —especialmente las otras— son ingenuas y creerán cualquier cosa que encuentren en Internet. Como demuestra Hugo Mercier (2020), la creencia en la ingenuidad humana está muy equivocada.

En relación con la tercera razón, Williams subraya que en las sociedades democráticas occidentales, el entorno mediático es altamente competitivo y las personas deben ser selectivas en cuanto a la información que buscan, debido a su limitado presupuesto de atención. La selección de fuentes a las que prestan atención se basa en la confianza y reputación de estas, así como si el contenido es entretenido o útil y si concuerda con su visión del mundo. Por lo tanto, productores

ISSN: 2952-6094 | e-ISSN: 2952-6116 | Núm. 12 | Julio de 2024
https://doi.org/10.62269/cavcaa.20 | Consejo Audiovisual de Andalucía

de contenido como compañías mediáticas, personalidades de Internet y teóricos de conspiraciones deben luchar por ganar la atención y confianza del público. Es sumamente difícil destacar, ya que la mayoría del contenido en plataformas como X o YouTube recibe poco o ningún *engagement* —como se dice en la jerga de los medios de comunicación—. En el contexto de la desinformación basada en IA, la cuestión es si puede superar a productores de contenido establecidos y fiables para ganar una audiencia de confianza. Además, dado que la desinformación es impulsada por la demanda y la mayoría se informa a través de fuentes establecidas, es improbable que la desinformación basada en IA capte una audiencia más amplia o tenga efectos significativos, ya que tiende a reforzar las creencias preexistentes más que a expandirlas.

En resumen, si bien la desinformación por IA es una preocupación, la competencia del mercado mediático y las preferencias ya establecidas del público crean obstáculos significativos para su influencia generalizada.

Finalmente, la cuarta razón: Williams sostiene que cuando la gente expresa preocupación sobre la desinformación, se refiere a mensajes que contradicen las narrativas del *establishment*, es decir, las visiones consensuadas de científicos y expertos, autoridades de salud pública, agencias de estadísticas aprobadas por el Gobierno, verificadores de hechos de medios de comunicación convencionales y élites de Davos, entre otros. ¿Cómo de preocupados deberíamos estar de que la desinformación basada en IA alimente el contenido contra-*establishment*? Pues resulta que muy poco. Porque, en las democracias occidentales, el *establishment*, es decir, gobiernos, medios de comunicación generalistas, organizaciones supranacionales (Europa), empresas legalmente dependientes de los contrapesos del Gobierno, etc. tienen mayor acceso a sistemas de IA más efectivos que las fuentes contra-*establishment*. En otras palabras y simplificando mucho, los «buenos» tienen más armas que los «malos».

En las secciones subsiguientes, me propongo desafiar la narrativa predominante que sugiere que los avances en IA, los cuales facilitan la creación de imágenes, vídeos y voces sintéticas, podrían amplificar la propaganda, la desinformación y los discursos de odio. Aunque existe cierto fundamento en esta preocupación, argumentaré que tales desarrollos tecnológicos no representan una amenaza significativa para la integridad de nuestras democracias.

Pero antes de ello me gustaría dejar claro que este escrito no es una alegato o apología de visiones tecnofílicas. La tendencia a adoptar visiones extremas —ya sean tecnofóbicas o tecnofílicas— puede entenderse como una manifestación del sesgo de confirmación, donde individuos y grupos tienden a favorecer información que confirma sus preconcepciones y a ignorar o desvalorizar la que contradice sus creencias.

En lugar de adoptar una postura unidimensional que asuma que la tecnología, por su propia naturaleza, es inherentemente buena o mala, es fundamental desarrollar un marco de análisis que reconozca la tecnología como una herramienta cuyo impacto depende del contexto social, político y económico en el que se implementa (Astobiza, 2023). Como dice el famoso adagio: «La tecnología no es buena ni mala, pero tampoco es neutra» (Kranzberg y Pursell, 1967).

1.1. Contexto actual: la democracia en la era digital

En el ámbito de la economía y las ciencias políticas, la interconexión entre democracia, tecnología y bienestar o crecimiento económico ha sido objeto de debate durante mucho tiempo.

Muchos expertos están de acuerdo en que los países ricos son democracias y que el desarrollo económico conduce a la libertad política y esto es gracias, en buena parte, al desarrollo tecnológico (Barro, 1999). La interconexión entre democracia, crecimiento, tecnología y el desempeño de los sectores industriales de un territorio determinado es fundamental.

La democracia es facilitadora del desarrollo de la tecnología y la tecnología es del crecimiento económico que a su vez refuerza los pilares de la democracia. La relación causal es evidente. La democracia garantiza las libertades políticas y permite la operación de organizaciones empresariales que innovan. Dicha innovación se traduce en tecnologías, servicios y productos que incrementan el bienestar de las personas. El mayor bienestar de las personas refuerza el deseo de tener las instituciones que se tienen, que resulta que acaba reforzando la democracia.

La tecnología afirma la democracia y la democracia afirma la tecnología. Las instituciones democráticas son más eficientes en la provisión de bienes públicos y servicios, lo que conduce a un mayor crecimiento económico (Acemoglu y Robinson, 2012), porque la tecnología es un factor clave en el crecimiento de la productividad (Solow, 1957).

La democracia puede ser vista como un facilitador del desarrollo tecnológico. La libertad de expresión, la participación política y la transparencia que caracterizan a las democracias crean un ambiente propicio para la innovación y el emprendimiento. Las instituciones democráticas también brindan estabilidad y previsibilidad, lo que reduce la incertidumbre y alienta la inversión en investigación y desarrollo.

A su vez, la tecnología juega un papel fundamental en el crecimiento económico. Las nuevas tecnologías generan nuevas ideas, productos y servicios que aumentan la productividad y la eficiencia. La tecnología también puede facilitar el acceso a la información y la educación, lo que a su vez puede mejorar las habilidades de la fuerza laboral y promover el desarrollo humano.

El crecimiento económico, por otro lado, puede fortalecer la democracia. Un mayor nivel de riqueza puede traducirse en una mejor calidad de vida para los ciudadanos, lo que puede aumentar su participación en la vida política y su confianza en las instituciones democráticas. El crecimiento económico también puede generar recursos adicionales para el Gobierno, que pueden ser utilizados para invertir en educación, salud y otras áreas clave para el desarrollo social.

Pero no cabe duda de que la tecnología por sí misma ha transformado profundamente al sociedad incluyendo la forma en la que ser ejerce la democracia. Internet y las redes sociales han abierto nuevas posibilidades para la participación ciudadana, la transparencia y el acceso a la información. Sin embargo, estas mismas tecnologías también han generado nuevos desafíos para la democracia, como la desinformación, la polarización y la manipulación electoral.

ISSN: 2952-6094 | e-ISSN: 2952-6116 | Núm. 12 | Julio de 2024
https://doi.org/10.62269/cavcaa.20 | Consejo Audiovisual de Andalucía

A lo largo de la historia, diferentes tecnologías han sido objeto de debate por su potencial impacto negativo en la esfera pública (Barber, 1984). La imprenta, la radio y la televisión fueron objeto de críticas por su capacidad para difundir ideas disidentes, propaganda o crear una cultura pasiva. En cada caso, los temores no se materializaron de la forma en que se predijo, pero sirven como recordatorio de que las nuevas tecnologías siempre tienen el potencial de ser utilizadas tanto para bien como para mal.

En el contexto digital, la desinformación, la polarización y la manipulación electoral son algunos de los principales desafíos para la democracia. La proliferación de información falsa y engañosa en Internet y las redes sociales puede erosionar la confianza en las instituciones democráticas y dificultar la toma de decisiones informadas por parte de los ciudadanos (Morozov, 2011). Las redes sociales también pueden crear cámaras de eco que aumentan la polarización política y dificultan el diálogo y la comprensión mutua (Sunstein, 2001). Además, las tecnologías digitales pueden ser utilizadas para manipular las elecciones, por ejemplo, mediante la compra de votos, *deepfakes* o la difusión de información falsa.

La era digital ha traído consigo una transformación radical en la forma en que consumimos y producimos información. El auge de Internet y las redes sociales ha democratizado el acceso a la información, pero también ha abierto las puertas a nuevos desafíos, como la desinformación y los *deepfakes*. No obstante, hemos visto con anterioridad que estos escenarios son menos probables de lo que parecen porque es muy difícil que las desinformación afecte nuestras creencias y mucho menos nuestro comportamiento.

Las personas son bastante buenas para detectar *deepfakes*, que son costosos y requieren de habilidades técnicas, y que suelen ser burdos y poco convincentes. Además, señalan que existe una creciente conciencia sobre la desinformación y los *deepfakes*, lo que hace que las personas sean más críticas con la información que encuentran en línea. Es importante no minimizar el potencial de daño que pueden causar la desinformación y los *deepfakes*. Sin embargo, tampoco debemos caer en la paranoia o alarmismo. Es necesario mantener una perspectiva equilibrada y considerar las siguientes medidas para mitigar sus riesgos:

- Educar a la población sobre la desinformación y los *deepfakes*. Es fundamental que las personas sean conscientes de estas tecnologías y aprendan a identificar contenido falso.
- Promover el pensamiento crítico y la verificación de información. Las personas deben ser críticas con la información que encuentran en línea y verificar su autenticidad antes de compartirla.
- Desarrollar herramientas para detectar *deepfakes*. La tecnología puede ser utilizada para combatir la desinformación y los *deepfakes* mediante el desarrollo de herramientas que puedan identificarlos con mayor precisión.

En definitiva, la tecnología tiene el potencial de ser utilizada para el bien o para el mal. Es nuestra responsabilidad asegurarnos de que se utilice para el bien de la democracia. Pero eso sí, la democracia en la era digital no se defiende sola, se defiende con participación activa, pensamiento crítico y responsabilidad.

2. La IA y la desinformación

La desinformación y los *deepfakes* son una amenaza real para la democracia. La capacidad de crear contenido falso y altamente convincente tiene el potencial de manipular a la opinión pública, influir en las elecciones y socavar la confianza en las instituciones democráticas.

Sin embargo, la desinformación basada en IA se puede evaluar empíricamente y combatirla. Se puede realizar investigación independiente con el fin de desarrollar estrategias que combatan la desinformación. Para ello es fundamental contar con datos y evidencia sólida para determinar qué enfoques funcionan mejor y desarrollar soluciones más efectivas.

La desinformación se refiere a la creación y difusión deliberada de información falsa y engañosa, generalmente con el objetivo de engañar al público, influir en la opinión pública o en los resultados políticos, o desacreditar a individuos o grupos (Wardle y Derakhshan, 2017). A diferencia de la información errónea, que puede ser compartida sin intención de engañar, la desinformación implica intencionalidad por parte de quien la produce.

Para combatir la desinformación es necesario la colaboración entre investigadores, tecnólogos, responsables políticos y la sociedad civil. La lucha contra la desinformación es un desafío complejo que requiere un esfuerzo multifacético y la participación de diversos actores. Para combatir de manera efectiva la desinformación y las *deepfakes* se necesita acceso abierto a los datos, herramientas y resultados de la investigación. La transparencia y la replicabilidad son esenciales para garantizar la confianza en la investigación y el desarrollo de soluciones sostenibles.

Es importante destacar que la desinformación basada en IA no es una fuerza imparable. Fortalecer la educación en medios audiovisuales y la alfabetización digital en todos los niveles de la sociedad, es una vía potencial y prometedora para conseguir combatir la desinformación basada en IA. Empoderar a los ciudadanos con las habilidades necesarias para identificar y cuestionar la desinformación es una parte esencial de la lucha contra este fenómeno. Además, es crucial promover un periodismo de calidad y fomentar un ecosistema mediático sólido y diverso.

La desinformación y las *deepfakes* impulsadas por la IA representan un desafío significativo para la democracia, pero no es una batalla perdida. A través de la investigación independiente, la colaboración multisectorial, el acceso abierto a los datos y herramientas, y la promoción de la alfabetización digital y audiovisual, es posible desarrollar estrategias efectivas para combatir este fenómeno y proteger la integridad de nuestros sistemas democráticos.

Una de las claves para abordar este reto es la inversión continua en investigación y desarrollo tecnológico. Los avances en áreas como la detección automática de *deepfakes*, la verificación de fuentes y la trazabilidad del contenido digital pueden proporcionar herramientas poderosas para identificar y contrarrestar la desinformación impulsada por IA. Además, es fundamental fortalecer los marcos regulatorios y legales para abordar este problema. Establecer normas y regulaciones claras sobre la creación y difusión de contenido falso o engañoso, así como sanciones para quienes lo hagan, puede disuadir a los actores maliciosos y promover un entorno digital más seguro y confiable.

ISSN: 2952-6094 | e-ISSN: 2952-6116 | Núm. 12 | Julio de 2024
https://doi.org/10.62269/cavcaa.20 | Consejo Audiovisual de Andalucía

La colaboración internacional también es clave. La desinformación basada en IA no conoce fronteras, por lo que es esencial que los gobiernos, las organizaciones internacionales y la sociedad civil trabajen juntos para compartir información, desarrollar estándares comunes y coordinar esfuerzos.

Pero más allá de los aspectos tecnológicos y regulatorios, es fundamental abordar las causas subyacentes que facilitan la propagación de la desinformación. Esto implica invertir en educación audiovisual, promover un periodismo sólido y confiable y fomentar una cultura de pensamiento crítico y verificación de hechos. No obstante, permítanme compartir un hallazgo crucial sobre la desinformación: es esencial desmontar los mitos prevalentes que la rodean. Principalmente, la preocupación por la desinformación impulsada por IA, como los *deepfakes*, suele ser exagerada.

En el debate contemporáneo sobre el estado de nuestra democracia y la salud de nuestro discurso público, la desinformación ocupa un lugar central, a menudo descrita como la amenaza definitiva a los cimientos democráticos. En el momento de redactar este artículo, en 2024, nos encontramos en lo que se denomina un «super año electoral», marcado por elecciones cruciales en India, Estados Unidos y Rusia (Mações, 2024). Y son muchos los académicos, científicos, periodistas, políticos e instituciones (Comisión Europea, 2022) que alertan de los peligros que la desinformación puede tener en la sociedad.

Pero una cosa no se nos debe olvidar. La desinformación, lejos de ser una novedad de la era digital, es un fenómeno tan antiguo como la propia sociedad. La noción de que nos encontramos en una crisis informativa única ignora la persistencia histórica de la propaganda, las teorías de conspiración sin fundamento y las ideologías inexactas como características omnipresentes en el tejido social humano.

Los defensores de la idea de una crisis sin precedentes suelen señalar a las redes sociales como el catalizador de esta era de desinformación, argumentando que, a diferencia de los medios tradicionales, que supuestamente se caracterizaban por su objetividad y veracidad, las redes sociales están inundadas de falsedades y teorías conspirativas. Sin embargo, esta comparación falla al ignorar las evidencias de parcialidad y desinformación que también han plagado a los medios tradicionales a lo largo de la historia.

Desde la cobertura sesgada de genocidios y hambrunas hasta la exageración de pruebas en conflictos armados, los medios tradicionales han estado lejos de ser baluartes de objetividad (Herman y Chomsky, 1988). Otro mito de la desinformación es que la mentira, los bulos y las falsedades se viralizan y propagan más rápidamente que la verdad. De hecho, un grupo de investigadores afirmaron esto en un trabajo que fue el estudio longitudinal con mayor muestra de noticias hasta la fecha (Vosoughi *et al*., 2018). No es mi cometido aquí hablar sobre la fiabilidad de este estudio, pero véase Aral (2022). Lo que sí podemos afirmar es que la desinformación no se propaga más rápidamente en las redes sociales que la información precisa porque la información siempre ha tenido la capacidad de alcanzar rápidamente a grandes audiencias, ya sea a través de periódicos, radio, televisión... De hecho, debido a la fragmentación de los medios audiovisuales actuales, la velocidad de difusión de los mensajes podría ser incluso más lenta hoy en día.

Los temores de que las redes sociales estén impulsando aumentos consistentes en malentendidos populares o en tendencias políticas preocupantes tampoco encuentran apoyo en la evidencia empírica. Estudios sugieren que las teorías conspirativas no son

más prevalentes que antes, la mayoría de los ciudadanos no están atrapados en cámaras de eco dañinas en línea, y los algoritmos de las redes sociales no están arrastrando a las personas hacia «agujeros de conejo» de desinformación (Nyhan, 2020).

Esto no implica que las redes sociales y otras plataformas alimentadas por algoritmos o tecnología basada en IA carezcan de efectos nocivos sobre las democracias, pero la narrativa es compleja, involucrando mezclas sutiles de costos y beneficios que varían según el contexto y son difíciles de cuantificar. La idea simplista de que las redes sociales, o, en términos generales, la IA y los *deepfakes* han inaugurado una crisis informativa sin precedentes carece de sustento empírico y argumentos persuasivos. En consecuencia, la solución a los desafíos democráticos que enfrentamos hoy en día requiere una comprensión más matizada y basada en evidencias del papel de la desinformación en la sociedad.

3. La IA y el discurso de odio

En un fallo significativo del 28 de agosto de 2018, el Tribunal Europeo de Derechos Humanos (TEDH) abordó el caso de Savva Terentyev contra Rusia, marcando un hito en la jurisprudencia sobre libertad de expresión (Savva Terentyev v. Rusia, 2018).

El TEDH dictaminó que Rusia había violado el derecho a la libertad de expresión de Terentyev, garantizado por el artículo 10 del Convenio Europeo de Derechos Humanos. Este veredicto es notable por varias razones, destacando especialmente la importancia del contexto al evaluar expresiones potencialmente ofensivas o insultantes y reafirmando la protección elevada de la libertad de expresión, especialmente en periodos electorales.

El caso se centró en un bloguero joven condenado a prisión por incitar al odio a través de comentarios despectivos sobre la policía, publicados en el marco de las acciones de las fuerzas de seguridad durante un proceso electoral. El TEDH subrayó que, aunque las palabras del bloguero eran duras y vulgares, se pronunciaron en el contexto de un debate sobre un tema de interés público, reiterando la importancia de permitir la circulación libre de opiniones e información antes de las elecciones.

El Tribunal también consideró que ciertas expresiones del bloguero, aunque crudas, poseían un sentido metafórico, sin intención de ofender a las verdaderas víctimas del Holocausto o incitar a la violencia. Se enfatizó que la policía, al ser una institución pública, está sujeta a un nivel de crítica mayor que los ciudadanos comunes y que debe demostrar una tolerancia elevada ante el discurso ofensivo, a menos que dicho discurso presente un riesgo real e inminente de provocar acciones ilegales.

Porque el discurso de odio debe ser evaluado atendiendo al contexto donde las palabras se enmarcan y a quién se dirige. El discurso de odio es cualquier forma de expresión que difunde, incita, promueve o justifica el odio racial, la xenofobia, el antisemitismo u otras formas de odio basadas en la intolerancia, incluyendo la intolerancia expresada mediante el nacionalismo agresivo y el etnocentrismo, la discriminación y la hostilidad contra minorías, migrantes y personas de origen inmigrante (Sellars, 2016). Este tipo de discurso es perjudicial porque perpetúa la estigmatización y la discriminación y puede incitar a la violencia, la exclusión o la represión. Por ello, si a quien se dirige es un grupo

ISSN: 2952-6094 | e-ISSN: 2952-6116 | Núm. 12 | Julio de 2024
https://doi.org/10.62269/cavcaa.20 | Consejo Audiovisual de Andalucía

vulnerable y hay riesgo de un peligro claro e inminente, entonces podemos hablar de discurso de odio. Aunque el discurso de odio es moralmente reprobable y puede incluso constituir un delito, demostrar que incita directamente a la violencia física presenta significativas dificultades.

Por eso muchos juristas son de la opinión de que restringir la libertad de expresión solo ha de hacerse cuando hay una necesidad clara y razones suficientes. Las palabras, por muy repugnantes que puedan ser, son objetos abstractos sin mucho poder causal. Muy rara vez las palabras llevan a actos de violencia. Cuando esto ocurre suelen ser una racionalización *post hoc* de actos que la gente ya iba a realizar de todas maneras.

Como apunta la literatura de investigación reseñada por Mercier (2020), la mayoría de la gente es relativamente racional y no se deja influir fácilmente por la publicidad, los eslóganes, las noticias falsas, bulos o desinformación, los hechos alternativos y, sobre todo, las redes sociales. Los partidarios de la censura deben demostrar que hay un problema que resolver.

Aun así no cabe duda que las redes sociales están plagadas de contenido antisemita, islamófobo, racista, sexista, misógino... Este contenido no debería preocuparnos tanto como las personas propensas a la violencia, reflejando la idea de que las palabras, incluido el discurso de odio, raramente ejercen una causalidad directa sobre actos violentos. El discurso de odio se puede instrumentalizar pero la predisposición violenta existe *ex ante*.

Por otra parte, la tecnología no es más culpable que las personas violentas aunque el contenido execrable se difunda por redes sociales. La tecnología, en sí misma, no debe ser señalada como culpable de propagar discursos de odio, ya que son las personas con predisposiciones violentas las que eligen utilizar estas herramientas para difundir mensajes perniciosos. Aunque las plataformas digitales pueden facilitar la velocidad y el alcance de la difusión de tales discursos, la raíz del problema reside en las intenciones y comportamientos de aquellos individuos inclinados a la violencia (Román San Miguel, Olivare García y Jiménez-Zafra, 2022).

La atención y los esfuerzos por mitigar el impacto del discurso de odio deberían centrarse más en abordar y reformar las actitudes y comportamientos de las personas propensas a la violencia, en lugar de atribuir la responsabilidad directamente a las herramientas tecnológicas que utilizan. Esta perspectiva invita a una reflexión más profunda sobre cómo las sociedades pueden prevenir la violencia y promover un diálogo más constructivo, respetando al mismo tiempo la libertad de expresión y el potencial positivo de la tecnología para el empoderamiento y la comunicación.

4. Conclusiones

Permítanme concluir con una cita que se atribuye a Marie Curie, una de las científicas más ilustres de la historia y distinguida con el Premio Nobel en dos ocasiones, en Física y en Química: «Nada en la vida debe ser temido, solamente comprendido. Ahora es el momento de comprender más para temer menos».

Con esta referencia, de la cual no tenemos ninguna certeza definitiva que confirme que ella la haya dicho o escrito, deseo destacar que, aunque los *deepfakes*, la desinformación

y los discursos de odio constituyen serios desafíos para la democracia, la tecnología no agudiza su gravedad ni los hace inevitables. Porque de la misma forma que no sabemos verificar la autoría de esa frase atribuida a Marie Curie pero dejamos constancia de ello, también podremos determinar si algo es real, *fake* o es inconcluyente poder determinar su genealogía. Es verdad que solo en el caso de los *deepfakes,* la ausencia de tecnología digital eliminaría por completo la posibilidad de crear imágenes, voces o vídeos sintéticos manipulados. No obstante, es importante reconocer que la tecnología también nos ofrece herramientas esenciales para identificar y contrarrestar estos *deepfakes*, desinformación y discursos de odio. Además, estos no representan un desafío mayor que nuestros propios sesgos internos, los cuales pueden llevarnos a desconfiar de información veraz y, a la inversa, a aceptar como ciertas afirmaciones falsas.

Las *deepfakes*, la desinformación y los discursos de odio, aunque preocupantes, no constituyen por sí solos una amenaza insuperable para la democracia. Reconocer esto no implica restar importancia a su impacto potencial. Es comprensible que nadie desee enfrentarse a vídeos alterados, voces fabricadas o la propagación de falsedades y mensajes racistas, islamófobos, misóginos, sexistas... en redes sociales o plataformas audiovisuales. La clave para mitigar su efecto nocivo yace en nuestra capacidad de entender cómo funcionan. Al profundizar en el entendimiento de estos fenómenos, disminuimos el poder que ejercen sobre nosotros y sobre la sociedad. Esta comprensión nos empodera para desarrollar estrategias efectivas, fomentando así una cultura de resiliencia informacional. Al enfrentarnos a ellos con conocimiento y herramientas adecuadas, no solo reducimos el miedo, sino que también reforzamos los cimientos democráticos frente a las distorsiones de la realidad.

Lo único que sí que puede ser peligroso es una tendencia que cada vez es más visible. Por ejemplo, se empieza a notar que la gente tiende a rechazar información que cree que ha sido generada por IA. Esta tendencia sí que es peligrosa, porque como hemos visto es muy poco probable que la IA genere contenido que dé lugar a burbujas de filtro o cámaras de eco. Los contenidos de IA no crearán cámaras de eco en las que la gente rechace lo que no cree. Pero es evidente que potenciará esta tendencia ya peligrosa, dando aún más excusas para no escuchar. En otras palabras, las personas podrían usar la generación de contenido por IA como pretexto para descartar, ignorar o negar información que no prefieran o no quieran aceptar.

La IA generativa implica riesgos notables no por el potencial de saturar el ciberespacio con deepfakes, desinformación o discursos de odio indistinguibles, sino por el peligro de que, ante la generalización de la percepción de que todo contenido es generado por IA, se desencadene una desconfianza masiva que entorpezca la formación de consenso. La proliferación de desinformación mediante *deepfakes* o discursos de odio podría llevar a un estado en el que solo se acepte como verdadero lo que ya se considera cierto, fomentando la desconfianza en todo lo demás. Este escenario podría conducir a una peligrosa anomia social.

Es fundamental destacar la interacción entre nuestras capacidades cognitivas y las tecnologías emergentes como la IA, especialmente en el contexto de los *deepfakes*, la desinformación y los discursos de odio en democracia. A menudo, nuestra percepción de la realidad está mediada no solo por nuestras experiencias directas, sino también por

ISSN: 2952-6094 | e-ISSN: 2952-6116 | Núm. 12 | Julio de 2024
https://doi.org/10.62269/cavcaa.20 | Consejo Audiovisual de Andalucía

las herramientas tecnológicas que utilizamos para interpretar y entender el mundo. Esta interacción puede llevar a una disonancia cognitiva cuando enfrentamos información que contradice nuestras creencias preexistentes o cuando interactuamos con medios que pueden ser manipulados.

Profundizar en nuestra comprensión de estos fenómenos tecnológicos nos permitirá no solo reducir el miedo y la desconfianza sino también mejorar nuestra capacidad para participar de manera informada y crítica en una sociedad democrática. A medida que la tecnología evoluciona, también debe hacerlo nuestra capacidad para comprender sus implicaciones. Esto no solo reforzará los cimientos democráticos frente a las distorsiones de la realidad, sino que también empoderará a las personas para enfrentar y desmantelar las estrategias de desinformación y odio, asegurando un espacio público más auténtico y menos polarizado.

Siguiendo la cita atribuida a Curie, es mejor entender que temer. No temamos el uso de la IA para la desinformación. Entendamos su verdadero alcance. Las personas no carecen de inteligencia; el desafío surge cuando el contenido generado por IA valida preconcepciones existentes.

Financiación

El presente trabajo no ha recibido ayudas específicas provenientes de agencias del sector público, sector comercial o entidades sin ánimo de lucro.

5. Referencias

Aral S. (2022, 6 de abril). *Fake News about our Fake News Study Spread Faster than its Truth… Just as We Predicted.* Disponible en: https://bit.ly/3TwE0Pt

Astobiza A. M. (2023). *Tecnofilosofía: Nuestra Relación con las Máquinas.* Madrid: Plaza y Valdés.

Acemoglu, D. y Robinson, J. A. (2012*). Why nations fail: The origins of power, prosperity, and poverty*. NY: Crown Business.

Barro, R. (1999). The Determinants of Democracy. *Journal of Political Economy* 107, S158-S183.

Bronstein, M. V., Pennycook, G., Bear, A. *et al*. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition,* 8(1), 108-117. Disponible en: https://doi.org/10.1016/j.jarmac.2018.09.005

Chesney, B. y Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review,* 107, 1753-1819.

Comisión Europea (2022). Tackling online disinformation. Disponible en: https://bit.ly/3v6TDDG

Herman, E. S. y Chomsky, N. (1988). *Manufacturing Consent: The Political Economy of the Mass Media.* Pantheon Book.

Kranzberg, M. y Pursell, C. W. Jr. (eds.). (1967). *Technology in Western Civilization,* vol. 1. Oxford: Oxford University Press.

McKay, R. y Dennett, D. (2009). The evolution of misbelief. *Behavioral and Brain Sciences, 32*(6), 493-510. Disponible en: https://doi.org/10.1017/S0140525X09990975

Mercier, H. y Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences,* 34(2), 57-74. Disponible en: https://doi.org/10.1017/S0140525X10000968

Mercier, H. y Sperber, D. (2017). *The Enigma of Reason*. Cam. Massachusetts: Harvard University Press.

Mercier, H. (2020). *Not Born Yesterday: The Science of Who We Trust and What We Believe.* Princeton, NJ: Princeton University Press.

Morozov, E. (2011). *The net delusion: The dark side of internet freedom.* PublicAffairs.

Mações B. (2024, 10 de enero). The year of voting dangerously. *The New Stetesman.* Disponible en: https://bit.ly/3v4fsUs

Nyhan, B. (2020). Facts and myths about misperceptions. *Journal of Ecnomics Pespectives,* 34 (3): 220-36.

Paris, B. y Donovan, J. (2019). Deepfakes and cheap fakes: The manipulation of audio and visual evidence. *Data & Society.* Disponible en: https://bit.ly/442FYdU

Román-San-Miguel, A., Olivares-García, F. J., y Jiménez-Zafra, S. M. (2022). El discurso de odio en Twitter durante la crisis migratoria de Ceuta en mayo de 2021. *Revista ICONO 14. Revista Científica de Comunicación y Tecnologías Emergentes,* 20(2). Disponible en: https://doi.org/10.7195/ri14.v20i2.1858

Saner E. (2024, 31 de enero). Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'. *The Guardian.* Disponible en: http://bit.ly/3v283F6

Sellars, A. (2016). Defining Hate Speech. *Berkman Klein Center for Internet & Society Research Publication*. Disponible en: https://bit.ly/4d1sl2A

Solow, R. M. (1957). Technical change and the aggregate production function. *Review of Economics and Statistics,* 39(3), 312-320.

Savva Terentyev v. Rusia, Aplicación No. 001-185307 (TEDH, 28 de agosto de 2018). Recuperado de http://hudoc.echr.coe.int/eng#{"itemid":["001-185307"]}

Sunstein, Cass R. (2001). *Republic.com 2.0: How the Internet is changing democracy.* Princeton University Press.

Vosoughi, S. *et al.* (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151. Disponible en: https://doi.org/10.1126/science.aap9559

Wardle, C. y Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe Report DGI(2017)09.* Disponible en: https://bit.ly/3Q5KKBs

Williams D. (2024, 24 de enero). AI-based disinformation is probably not a major threat to democracy. Disponible en: https://bit.ly/3Isv8Us

ISSN: 2952-6094 | e-ISSN: 2952-6116 | Núm. 12 | Julio de 2024
https://doi.org/10.62269/cavcaa.20 | Consejo Audiovisual de Andalucía