Revista de Ciencias de la Comunicación e Información. 2024. Vol. 29, 1-14
ISSN 2695-5016

VÍDEOS FALSOS Y DESINFORMACIÓN ANTE LA IA: EL DEEPFAKE COMO VEHÍCULO DE LA POSVERDAD

FAKE VIDEOS AND DISINFORMATION BEFORE THE IA: DEEPFAKE AS A POST-TRUTH VEHICLE

Lucia Ballesteros Aguayo[1]**:** *Universidad de Málaga. España*.

Francisco Javier Ruiz del Olmo: *Universidad de Málaga. España*.

Financiación. *El presente texto nace en el marco del proyecto Posverdad a debate: reconstrucción social tras la pandemia (P020_00703).*

Cómo citar el artículo:

Ballesteros Aguayo, Lucia y Ruiz del Olmo, Francisco Javier (2024). Vídeos falsos y desinformación ante la IA: el *deepfake* como vehículo de la posverdad [Fake videos and disinformation before the AI: deepfake as a post-truth vehicle]. *Revista de Ciencias de la Comunicación e Información*, 29, 1-14. https://doi.org/10.35742/rcci.2024.29.e294

RESUMEN

Introducción: El uso de la Inteligencia Artificial en la generación de contenido y narraciones audiovisuales si bien representa una oportunidad en muchos campos como el artístico o en la creación visual y gráfica, también se convierte en un potente instrumento para generar relatos y representaciones falsos. **Metodología:** Se aplica la Revisión Sistemática Exploratoria (RSE), aportando referencias que radiografíen con evidencias empíricas la imagen de la posverdad. **Resultados:** Se aporta una revisión crítica de los últimos estudios y tendencias en la creación de imagen mediante inteligencia artificial relacionadas con la desinformación. Ésta forma parte del ecosistema audiovisual contemporáneo amenazando la confianza de la ciudadanía en el entorno mediático, social o institucional. **Discusión:** Los usuarios, a través de las redes sociales, generan imágenes falsas o distorsionadas, que una vez viralizadas son nuevamente reinterpretadas por otros usuarios. Los vídeos falsos pueden arruinar tanto la reputación del individuo como la confianza en los actores sociales. Estos efectos podrían estar moderados por la alfabetización visual y digital. **Conclusiones:** El aprendizaje profundo de las redes neuronales artificiales genera nuevas formas de *deepfake*, desconcertantes por su realismo y verosimilitud, y que empiezan a suponer un cuestionamiento hacia los medios de comunicación, deslegitimando la representación de la realidad y la información veraz como base de una sociedad democrática.

Palabras clave:

Desinformación; Inteligencia Artificial; *Deepfake*; Posverdad.

[1] **Lucia Ballesteros Aguayo:** Profesora en la Facultad de Comunicación de la Universidad de Málaga. Líneas de investigación: *Communication*, *Disinformation*, *Propaganda* y *Media Studies*.

Recibido: 27/12/2023 - Aceptado: 12/03/2024 - Publicado: 04/04/2024

ABSTRACT

Introduction: The use of Artificial Intelligence in the generation of audiovisual content and narratives, although it represents an opportunity in many fields such as art or visual and graphic creation, also becomes a powerful tool to generate false stories and representations. **Methodology:** The Exploratory Systematic Review (ESR) is applied, providing references that show with empirical evidence the image of post-truth. **Results:** A critical review of the latest studies and trends in image creation through artificial intelligence related to disinformation is provided. This is part of the contemporary audiovisual ecosystem, threatening citizens' trust in the media, social or institutional environment. **Discussion:** Users, through social networks, generate false or distorted images, which once viralized are again reinterpreted by other users. Fake videos can ruin both an individual's reputation and trust in social actors. These effects could be moderated by visual and digital literacy. **Conclusions:** The deep learning of artificial neural networks generates new forms of deepfake, disconcerting for their realism and verisimilitude, and which begin to question the media, delegitimizing the representation of reality and truthful information as the basis of a democratic society.

Keywords:

Disinformation; Artificial Intelligence; Deepfake; Post-truth.

1. INTRODUCCIÓN

El consumo intensivo y acelerado de contenidos audiovisuales tanto de carácter informativo como de entretenimiento en las redes sociales, unido al desarrollo de una potente tecnología basada en la lógica algorítmica y el aprendizaje automático, están transformando nuestra forma de acercarnos a la realidad, y a la verdad.

La cada vez más acuciante viralización de relatos y representaciones falsos con escaso control por parte de autoridades, instituciones, ciudadanos y la insuficiente regulación por las propias plataformas —vehículo de dichos productos falsos—, es objeto de vivos debates en el seno de diversos organismos supranacionales como la Comisión Europea, entre otros. En diciembre de 2023, el Consejo y el Parlamento Europeo alcanzaron un acuerdo provisional sobre la que será la primera Ley de Inteligencia Artificial en el marco europeo que garantice una IA segura y transparente. El objetivo último es promover la adopción de una IA confiable y centrada en el ser humano y proteger la salud, la seguridad, los derechos fundamentales y la democracia de sus efectos nocivos (European Parliament, 2023). Entre las prácticas prohibidas por parte de la IA por su uso intrusivo y discriminatorio, los eurodiputados tipificaron la extracción no selectiva de imágenes faciales de Internet o de imágenes de circuito cerrado de televisión (CCTV) para crear bases de datos de reconocimiento facial. Las sanciones por incumplimiento de la citada ley oscilen entre 35 millones de euros o el 7% de la facturación global y 7,5 millones de multa.

Especial relevancia comporta el uso de la IA en la difusión de contenido a través de las redes sociales y más concretamente en su aplicación sobre los materiales audiovisuales; si bien esta tecnología emergente representa una oportunidad en muchos campos como el artístico o la creación audiovisual y gráfica, también supone un desafío para la ciudadanía a la hora de distinguir entre lo real y lo ficticio, entre la verdad y la mentira, esto es, se convierte en un potente instrumento para generar relatos y representaciones falsos. Autores como Shahid *et al*. (2022) advierten de que la mayoría de los usuarios carecen de las habilidades y la voluntad para detectar vídeos falsos y no son conscientes de los riesgos

y daños que entraña este tipo de falsedad. Expresamente en países como India -donde estos investigadores centran su estudio- destaca el hecho de que incluso cuando los usuarios saben que se trata de un vídeo falso, prefieren no realizar ninguna acción y, a veces, comparten voluntariamente vídeos no veraces pero que favorecen su visión del mundo.

Y es que desde hace más de una década, se implementa un giro de tendencia en el uso y consumo de los discursos en línea que priorizan o descansan en el discurso audiovisual. La popularidad de las plataformas de vídeo como TikTok, Instagram o YouTube está estimulando el consumo de vídeo a expensas del texto. De forma que el auge del formato vídeo es no sólo imparable sino permanente y consolidado. Así lo evidencia el Digital News Report (Neuman *et al*., 2023) que pone de manifiesto cómo en 2023 se produjo una paulatina caída en la participación en plataformas tradicionales como Facebook, mientras que se incrementaba en TikTok y otras redes basadas casi totalmente en el formato vídeo. También el consumo de noticias en vídeo está experimentando un crecimiento absoluto en todos los mercados.

Esto es aún más relevante si tenemos en cuenta que los jóvenes consumen desproporcionadamente más vídeos de noticias en las redes sociales al mismo tiempo que es menos probable que accedan a vídeos en los sitios web o aplicaciones de específicos de noticias. Concretamente, el informe de Neuman *et al*. (2023) subraya que los usuarios de entre 18 y 24 años se inclinan por el consumo de los vídeos cortos de TikTok, los *Reels* de Instagram y los YouTube *Shorts*. Por todo ello y ante estas nuevas prácticas de la posverdad y la desinformación con imágenes falsas creadas mediante IA, este trabajo realiza un estudio exploratorio de las principales tendencias y estudios que abordan este fenómeno como base para futuras preguntas y procesos de investigación.

En este tejido de desórdenes informativos (Wardle y Derakhshan, 2017), el Digital News Report de 2020 (Neuman *et al*., 2020) ya advertía de que las redes sociales eran para los usuarios una mayor fuente de preocupación sobre la difusión de desinformación (40%), por delante de los sitios de noticias (20%), aplicaciones de mensajería como WhatsApp (14%), y buscadores como Google (10%). Resulta particularmente grave la utilización en estas plataformas online del *deepfake* por su capacidad para generar desinformación de forma muy poderosa y con gran verosimilitud (Langguth *et al.,* 2021).

Los vídeos falseados o *deepfakes* son falsificaciones audiovisuales creadas deliberadamente para sugerir que alguien hizo o dijo algo que nunca ocurrió (Chesney y Citron, 2018; Nelson y Lewis, 2019). Cabe destacar el salto exponencial que han supuesto estos mecanismos de manipulación por su capacidad para distorsionar la realidad de forma espectacular e impactante. A ello se une la posibilidad de difusión rápida y generalizada y el hecho de que puedan ser utilizados por parte de usuarios que no necesitan poseer un dominio amplio de la tecnología. El resultado son falsificaciones profundas cada vez más realistas y resistentes a la detección.

2. OBJETIVOS

La desinformación tiene en la actualidad un poderoso aliado en la preponderancia de la imagen y de los discursos audiovisuales, tanto en la transmisión de los contenidos informativos en los diversos medios como por la influencia de las redes sociales, cuya relación con la realidad es cada vez más problemática, siendo alterada o transformada.

El presente texto persigue, mediante el método de revisión exploratoria y descriptiva, un objetivo principal que se concreta en tres objetivos específicos; como objetivo principal se busca evidenciar el poder de la imagen como elemento referencial de la realidad en los procesos de desinformación, señalando las principales aportaciones críticas de esta tendencia comunicativa a partir del papel de la literatura existente sobre el rol de la imagen creada por la inteligencia artificial, los vídeos falsos y los *deepfakes*, como elementos que alteran sustancialmente la realidad. Como objetivos secundarios, que se extraen del análisis crítico exploratorio de textos científicos relevantes publicados sobre la materia, se encuentran:

- Señalar las características de la tecnología emergente basada en la inteligencia artificial y el aprendizaje automático y su capacidad para falsear la realidad visualmente.
- Subrayar la influencia de los vídeos falseados o *deepfake* como elemento de materialización y amplificación de la posverdad, especialmente referida a la generación millennial.
- Apuntar los principales riesgos y soluciones al fenómeno de las falsificaciones mediante software de generación de vídeos falsos.

3. METODOLOGÍA

Siendo la generación de imágenes para la desinformación mediante inteligencia artificial un fenómeno relativamente reciente y que se empieza en la actualidad a generalizar, en este trabajo se ha optado, en conexión con los objetivos propuestos, por una metodología que descansa en la Revisión Sistemática Exploratoria (RSE), siguiendo a Booth *et al*. (2012) y a Munn *et al*. (2018). Se trata de un tipo de revisión crítica de estudios, fuentes y referencias que permiten explorar las tendencias en la reflexión sobre la temática propuesta y sintetizar evidencias empíricas, en este caso sobre la imagen de la posverdad. Con ello no solo se aporta a la comunidad científica la crítica de un fenómeno emergente, sino que su finalidad es también generar nuevas líneas de investigación.

Uno de los objetivos de la RSE es mostrar un panorama más amplio y contextualizado del fenómeno estudiado, lo que difiere de una revisión sistemática tradicional, más específica y exhaustiva. La RSE se aplica en varias etapas que incluyen la elaboración de la pregunta de investigación y formulación de una estrategia de búsqueda, la búsqueda de la literatura, la revisión y selección de los estudios, la extracción de datos y tendencias y el análisis e informe de resultados. En esta investigación se ha optado por una RSE que muestre un panorama amplio, mejor que exhaustivo, mediante "aproximaciones sistemáticas" (Booth *et al*., 2012) para dar respuesta a las tendencias e inquietudes que plantea difusión de imágenes de la IA.

Para el presente análisis se realizó una revisión exploratoria de literatura científica desde 2017, año de inicio en la difusión y repercusión masiva de vídeos falsos a través de redes sociales. Se buscó tanto artículos científicos, de tipo aplicado, de diseño experimental o no experimental, como artículos de discusión teórica, y también una selección de monografías en las bases de datos Scopus, Proquest, Ebsco y Dialnet. Para procesar la búsqueda se empleó el operador boleano "*and*" en los idiomas inglés y español. Los términos utilizados fueron "Vídeos falsos", "*Deepfake*", "Inteligencia Artificial". Se aplicó una delimitación temporal entre los años 2017 y 2023 y una delimitación geográfica referida a Europa, Latinoamérica y España.

La mayoría de los documentos se muestran lógicamente en varias bases de datos a la vez; el proceso de selección partió de la lectura y el análisis tanto del título como del abstract del documento; y en cuanto al criterio de selección, es necesario señalar que se trata de una revisión crítica y exploratoria, por lo que no se busca mostrar la totalidad de los resultados, sino que se hace una selección cualitativa e interpretativa de aquellos que integren los términos de búsqueda y ofrezcan aportaciones relevantes a los objetivos específicos planteados. Así es como diversas aportaciones científicas de estos textos se señalan e integran en el cuerpo de este trabajo.

4. RESULTADOS Y DISCUSIÓN

La imagen falseada, descontextualizada o reelaborada forma parte casi ineludible del ecosistema audiovisual contemporáneo. Con motivaciones y propósitos diferentes, estas imágenes extienden una permanente sensación de desconfianza hacia los medios, incluidos los especializados en información. Ciertamente en ocasiones estos mismos medios han actuado de forma poco ética, pero en todo caso la desinformación visual en la actualidad, al alcance de cualquier usuario, y la viralización de los mensajes pueden elevar este problema a un nivel estructural y de amenaza social o institucional. Por ejemplo, la proliferación de vídeos falsos relacionados con la actuación de las instituciones y sus representantes democráticos es uno de los principales desafíos de esta tecnología (Dan *et al*., 2021).

De todas las amenazas relatadas en la literatura, dos categorías amplias parecen particularmente relevantes, como han establecido Chesney y Citron (2018) o Vaccari y Chadwick (2020), entre otros. En primer lugar, los vídeos falsos pueden arruinar la reputación del individuo involucrado. Por ejemplo, en el caso de un candidato político, un vídeo falso podría afectar las actitudes del público y amenazar el éxito político del individuo. En segundo lugar, pueden surgir efectos indirectos perjudiciales, como una desconfianza generalizada en los actores sociales y políticos y una sensación de irrealidad y confusión sobre lo que es real y lo que no lo es. Dichos efectos podrían estar mediados por el realismo percibido (subjetivo) y moderados por la alfabetización visual y digital, entre otros.

A continuación presentamos algunas de las definiciones sobre *deepfake* más clarificadoras dada la escasa literatura al respecto. También se seleccionan aquellas aportaciones que inciden especialmente en los desafíos y retos que comporta el uso deshonesto de esta tecnología.

Tabla 1. Definiciones de *deepfake.*

Cover, R. (2022). *Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation*.	Los *deepfakes* recurren a los poderes algorítmicos, el aprendizaje automático y las modernas capacidades de procesamiento de la información para permitir a los usuarios insertar el rostro, el cuerpo y la información visual de una persona del mundo real en un escenario falso, produciendo vídeos muy convincentes que parecen ser un registro "verdadero".

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P. y Schroeder, D. T. (2021). *Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes.*	Es una tecnología novedosa que permite la manipulación económica de material de vídeo mediante el uso de inteligencia artificial.
Vaccari, C. y Chadwick, A. (2020). *Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News*.	Los *deepfake* son vídeos sintéticos que se asemejan mucho a los vídeos reales. Integrando teorías sobre el poder de la comunicación visual y el papel que desempeña la incertidumbre a la hora de socavar la confianza en el discurso público.
Nelson, A. y Lewis, J. A. (2019). *Trust your eyes? Deepfakes policy brief.*	Los *deepfakes* son falsificaciones de vídeo y audio casi perfectas producidas por programas de inteligencia artificial que producen imágenes y sonidos aparentemente realistas pero fabricados que muestran a personas haciendo y diciendo cosas que nunca sucedieron
Chesney, R. y Citron, D. K. (2018). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.	Las técnicas de aprendizaje automático están aumentando la sofisticación de la tecnología, haciendo que las falsificaciones profundas sean cada vez más realistas y cada vez más resistentes a la detección.

Fuente: Elaboración propia.

Si bien en países como España los escándalos más destacables de vídeos y/o imágenes falsificados mediante inteligencia artificial no aparecen hasta el año 2023 —es el caso por ejemplo de los menores en Almendralejo (Badajoz) que difundieron fotografías falsas de desnudos de chicas adolescentes, el falso desnudo de la cantante Rosalía o el de la modelo Laura Escanes—, la tecnología *deepfake* tiene su origen en 2017 cuando aparecieron vídeos en Reddit con rostros de mujeres famosas como Gal Gadot o Scarlett Johanssen injertados en cuerpos de actores porno (Cole, 2017; Langguth *et al*., 2021).

De esta manera y a partir del año 2017, el *software* de manipulación de imágenes para consumidores que se servía del aprendizaje automático ganó la atención del público (Dan *et al*., 2021). Sin embargo, en un artículo publicado cinco años antes, Krizhevsky *et al*. (2012) evidenciaron cómo el aprendizaje profundo, un perfeccionamiento de las redes neuronales artificiales, se estableció como una tecnología superior para el reconocimiento de imágenes. Para Langguth *et al*. (2021) dicho artículo demostró que cuando se entrenan con un gran número de imágenes de entrada adecuadas, las redes neuronales convolucionales (CNN) pueden categorizar el contenido de una imagen con gran precisión. De modo que una CNN puede entrenarse para que reconozca a personas concretas y distinguirlas con fiabilidad en una amplia gama de imágenes. Los requisitos para ello son un ordenador potente y un gran número de imágenes de las que la CNN pueda aprender. El

resultado —indicen estos autores— es una imagen creada a partir de las características abstractas que la red ha aprendido.

Por tanto, si bien el *software* es relativamente accesible y apenas requiere sofisticación, necesita de una gran cantidad de imágenes de entrenamiento para funcionar correctamente. En consecuencia, advierten Langguth *et al*. (2021), estos programas no son prácticos para crear vídeos manipulados de una persona promedio. Por esa razón, la mayoría de los vídeos *deepfake* que se elaboran con fines de entretenimiento presentan actores famosos de los cuales hay muchas imágenes disponibles públicamente.

Alineado con esta disponibilidad de las herramientas de manipulación de imágenes a través de la IA surge el término *cheapfake*. Gamir-Ríos y Tarullo (2022, p. 102) observan que "la tosquedad de su elaboración las hace más identificables como fraudulentas, las cheapfakes pueden producirse sin necesidad de habilidades tecnológicas avanzadas ni de sofisticados programas" y logran un efecto similar de polución desinformativa (Dowling, 2021), confirmando juicios preexistentes (Weeks y Garrett, 2014).

Lo cierto es que la manipulación de fotografías y vídeos tiene una larga historia, pero sólo los recientes avances tecnológicos han socavado la confiabilidad de las pruebas en la imagen como elemento referencial de la realidad.

4.1. Formas audiovisuales de la posverdad

Un exponente de este tipo de vídeos en línea es la publicación en otoño de 2018 en el medio estadounidense de entretenimiento BuzzFeed en la que aparecía Obama afirmando que "El presidente Trump es un total y completo imbécil" doblado por el actor Jordan Peele con el título "Tú ¡No creeré lo que dice Obama en este video! 😉 " que cosechó más de 5 millones de visitas en YouTube.

En el contexto del consumo mediático actual, estas prácticas no solo son comunes, sino que cada vez están más elaboradas; proliferan experiencias vicarias mediadas por su representación a través principalmente de vídeos e imágenes, de naturaleza fragmentaria, subjetiva y emocional, que niegan el carácter absoluto del conocimiento. En el sistema mediático contemporáneo, la posverdad sería de este modo una "inflación mediática" de lo posmoderno (Ferraris, 2019).

Las operaciones culturales de la posverdad y la imagen usada como su principal vehículo suelen estar informadas por una perspectiva parcial, sesgada, interesada y relativista (Prozorov, 2019). De manera que todo lo que antes se vivía directamente, ahora se ha trasladado a una representación (Debord, 1967). Esto es precisamente lo que sostiene el concepto de imagen notarial: parece irrefutable cuando nos lo muestran en una fotografía (Sontag, 1981).

En relación con quién y cuándo desinforma, destacamos tres grandes grupos o tipos de desinformación, esto es, la manera en que se concreta la posverdad en los discursos de los medios:

- Discursos que cuestionan la realidad a través de elementos satíricos o paródicos: las caricaturas y los memes con todos sus derivados.
- De naturaleza oportunista o inmediata: aferrados a la rapidez y vertiginosidad

de las noticias y acontecimientos narrados por los medios.
- Desinformación maliciosa: aquella que busca distorsionar de manera radical la percepción pública sobre una cuestión política, social o científica, entre otras.

Se observa que las manipulaciones de la imagen a través de los *deepfake* o de la inteligencia artificial progresivamente se ubican en el último grupo, dado que trabajan con manipulaciones profundas y altamente distorsionadores de la realidad.

4.2. La imagen: vehículo de la posverdad

Las prácticas de la posverdad potencian las características antes apuntadas a través de la imagen fija o de la narración audiovisual. Esto se debe al carácter polisémico de la imagen: es profundamente ambivalente respecto a su relación con la realidad. De ahí que subraye su capacidad para disfrazarse o aparentar similitud con la verdad. A ello se suma el carácter emocional e instantáneo, de modo que se trata de un formato que multiplica las capacidades distorsionadoras de la posverdad: la primacía de lo emocional frente a lo racional (Hameleers *et al*., 2020).

El significado de las imágenes no tiene por qué estar relacionado con el dispositivo técnico, con la tecnología asociada a la captación de las imágenes, esto es, no es resultado automático de la materialidad de la imagen, según Comolli (2010), tal y como detalla en su ensayo sobre Técnica e Ideología, sino con la interpretación de la misma. Así, la imagen refleja e informa directamente sobre la realidad representada, pero a la vez tiene una poderosa capacidad para la evasión, la subjetividad y la interpretación, combustibles impulsores del falseamiento de esa realidad. Por ejemplo, con relación a las imágenes que se basan en contenido falso o distorsionado se distinguen tres categorías generales:

- Las imágenes de descontextualización.
- Los memes.
- Las imágenes manipuladas: se corresponden con operaciones de alta sofisticación y profunda transformación, que se generan a través de la inteligencia artificial u otras técnicas de alteración radical.

Además, y como otro elemento clave que consolida la validación de las falsificaciones de las imágenes, los relatos de la posverdad mediática tienden a validar y reafirmar nuestras opiniones preexistentes, permitiéndonos ajustar los hechos a nuestro sistema de creencias personal (Lynch, 2016). Así, siguiendo a Lilleker y Liefbroer (2018), las emociones y las opiniones que parecen creíbles son más poderosas que las razonadas y verificadas. En este contexto cabe señalar la preeminencia cognitiva de las emociones y la intuición; el pensamiento automático, no reflexivo, y no crítico; la tendencia a las decisiones rápidas y reactivas ante estímulos espontáneos; y el origen de un juicio sesgado o estratégico (Lewandowsky *et al*., 2017; McDermott, 2019).

Por último, la reflexión crítica en torno a las imágenes de la posverdad, su naturaleza y sentido, ha sido acometida de forma histórica y pionera por dos instancias, antes de que llegara a ser un problema social e institucional importante y sea en la actualidad abordado ya de forma más sistemática por la comunidad académica y también por la profesión periodística:

A. La digitalización de la imagen y su transformación interesada, y la fotografía analógica, es también una representación y una manipulación (Marzal-Felici, 2021)

B. La generación intencionada de imágenes falsas y la reflexión: cuestiona la relación de las imágenes con la realidad (arte político) cuyo objetivo principal no es desarrollar narraciones falsas, sino permitir que los espectadores desarrollen una actitud crítica hacia las imágenes que consumen (Fontcuberta, 2017).

4.3. Imágenes falsas creadas con redes neuronales y su verificación

Todo proceso comunicativo precisa de una base tecnológica para llevarse a cabo, por su inherente carácter de representación y también por su necesaria transmisión y difusión. Los discursos posverdaderos aumentan y se viralizan gracias a la digitalización. La tecnología permite la creación de contenido prácticamente autónomo -tanto en formato texto como audiovisual- a través de la inteligencia artificial y de las redes neuronales que genera alteraciones profundas en la naturaleza y el sentido del discurso, siendo en la actualidad herramientas generadoras de las imágenes falsas. El desarrollo de este tipo de herramientas sostenidas por la IA constituye un hito en la cultura visual contemporánea, pues produce cambios sustanciales en la significación de la imagen, su generación y distribución.

Además, el ecosistema mediático actual se define por la intervención de las audiencias en los procesos productivos, en la oferta digital especializada, potenciando la interactividad en la comunicación, el consumo multipantalla y las narraciones trasmedia. De modo que a menudo son los usuarios, con o sin interés malévolo, los que generan imágenes falsas o distorsionadas, que una vez viralizadas son nuevamente reinterpretadas por otros usuarios.

Las redes neuronales que aplican la inteligencia artificial u otras herramientas digitales para transformar y falsear los relatos, noticias e informaciones, son a la vez generadores y verificadores de estos discursos. Las compañías que poseen y gestionan esta tecnología, conscientes de los desafíos y potenciales peligros y otras consideraciones éticas que su uso conlleva, han puesto en marcha diversos sistemas de autocontrol y verificación.

Algunos ejemplos recientes sobre esta cuestión los encontramos en el caso del software ChatGPT, que aunque no es capaz de discernir si la información es verdadera o falsa, puede proporcionar información adicional que permitirá al usuario obtener criterios sólidos externos para su verificación. Por tanto, mediante ChatGPT no es posible crear una *fake news* de forma directa. Ante la petición de crear una información falsa o crear una noticia poco contrastada o falta de argumento, la IA no genera texto, sino que advierte al usuario de que no puede crear dicha información por sus protocolos. La IA posee entonces herramientas internas para identificar las *fake news* y responde a los parámetros básicos de alfabetización mediática (Dan *et al*., 2021; Qian *et al*., 2022).

Otro ejemplo reciente lo encontramos en el gigante tecnológico Intel que ofrece *Real-Time Deepfake Detector*, una plataforma para analizar vídeos, reafirmando en el concepto de IA responsable. La herramienta, una de las primeras plataformas de detección de *deepfakes* en tiempo real del mundo, ofrece una alta tasa de precisión y está implementada en varios sectores y plataformas, incluidas herramientas de redes sociales, agencias de noticias, emisoras, herramientas de creación de contenido, nuevas empresas y organizaciones sin fines de lucro (Intel, 2022). Este detector de *deepfake* se basa en el estudio de los cambios de color en las caras al inferir el flujo de la sangre, un proceso llamado fotopletismografia.

Los investigadores Ilke Demir y Umur Ciftci diseñaron el software para centrarse en ciertos patrones de color en determinadas áreas faciales e ignorar otros, pudiéndose verificar la realidad o la falsificación de las imágenes.

5. CONCLUSIONES

El uso, la distribución y el consumo masivo, acelerado y persistente de imágenes en el ecosistema comunicativo contemporáneo conlleva retos sustanciales para nuestras sociedades. La necesidad de un marco normativo que regule los usos y límites de la IA constituye una de las prioridades de organismos supranacionales como la Unión Europea.

Las capacidades polisémicas de la imagen, su fijación emocional y representativa le confiere un poder central en la materialización y amplificación de la posverdad, de ahí los efectos nocivos y multiplicadores que conlleva la falsificación de imágenes y vídeos mediante *deepfake*.

La inteligencia artificial puede con un nivel de realismo creciente, construir, imaginar y transformar los sucesos y las narrativas de la actualidad para subvertirlas de forma profunda. Urge una reflexión crítica y la aportación de soluciones que permitan prevenir las consecuencias nefastas de la utilización de esta tecnología con fines equivocados, como puede ser el fraude electoral, el uso indebido difamatorio, el ciberbulling o la pornografización de figuras públicas o ciudadanos anónimos (Cover, 2022; Langguth *et al*., 2021).

Junto a ello, la viralización de vídeos falsos hiperrealistas que utilizan intercambios de caras posee un fuerte impacto en la credibilidad que tienen las personas hacia las pruebas audiovisuales en el periodismo (Shin y Lee, 2022). De ahí que los periodistas y los medios de comunicación estén claramente preocupados y sientan la responsabilidad de desacreditar estos vídeos falsos y evitar la manipulación de la opinión pública (Pérez Dasilva *et al*., 2021). Esta última investigación centrada en el uso de *deepfake* en Twitter destaca cómo en todos los tuits analizados, los medios informativos alertan del peligro potencial de esta tecnología. Esto coincide con estudios recientes como los aportados por Yadlin-Segal y Oppenheim (2020) que muestran cómo los periodistas enmarcan los *deepfakes* como una plataforma desestabilizadora que socava un sentido compartido de sociabilidad y de realidad política.

La literatura científica (Chesney y Citron, 2018; Vaccari y Chadwick, 2020) evidencia que las falsificaciones audiovisuales con mayor impacto son, por una parte, vídeos políticos sintéticos que generan el engaño y que se destinan a socavar la reputación del político o personalidad pública involucrada. Por otra parte, también pueden surgir efectos indirectos perjudiciales, como la desconfianza en los actores sociales y políticos y una sensación generalizada de confusión sobre lo que es real y lo que no lo es. Todo ello invita a reevaluar la ética cultural de la comunicación, más aún cuando una cantidad creciente de herramientas de *software* están hoy día a disposición de cualquier usuario, lo cual permite crear imágenes sintéticas desligadas del referente físico.

El uso generalizado de imágenes falsificadas en la desinformación supone un riesgo adicional para la credibilidad de las instituciones y constituye un desafío económico y creativo para las sociedades contemporáneas. Por el contrario, el volumen y la velocidad de la desinformación que prolifera a través de las plataformas en línea son una fuente de enormes ganancias para esas plataformas en línea (Dan *et al*., 2021). Las operaciones

culturales de la posverdad y la imagen usada como su principal vehículo suelen estar informadas por una perspectiva parcial, sesgada, interesada y relativista (Prozorov, 2019).

Entre las soluciones, los efectos de la creación de representaciones falsas mediante la IA podrían estar mediados por el realismo percibido y moderados por la alfabetización visual y digital (Dan *et al*., 2021; Qian *et al*., 2022). De modo que la intervención de alfabetización mediática digital puede motivar la búsqueda inversa de información visual errónea fuera de contexto. Otras soluciones (Chesney y Citron, 2018) se dirigen a la implicación de la industria con respuestas tecnológicas, sanciones penales, responsabilidad civil o acciones regulatorias.

6. REFERENCIAS

Booth, A., Papaioannou, D. y Sutton, A. (2012). *Systematic approaches to a successful literature review*. Sage.

Chesney, R. y Citron, D. K. (2018). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3213954

Cole, S. (2017). *AI-assisted fake porn is here and we're all fucked. Motherboard*. VICE. https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn

Comolli, J. L. (2010). *Cine contra espectáculo seguido de Técnica e ideología: 1971-1972.* Manantial.

Cover, R. (2022). Deepfake culture: The emergence of audio-video deception as an object of social anxiety and regulation. *Continuum*, *36*(4), 609-621. https://doi.org/10.1080/10304312.2022.2084039

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., Van Der Linden, S. y Von Sikorski, C. (2021). Visual Mis- and Disinformation, Social Media, and Democracy. *Journalism & Mass Communication Quarterly*, *98*(3), 641-664. https://doi.org/10.1177/10776990211035395

Debord, G. (1967). *The society of the spectacle*. Zone Books.

Dowling, M. E. (2021). Democracy under siege: Foreign interference in a digital era. En *Australian Journal of International Affairs*, *75*(4). Australian Institute of International Affairs (pp. 383-387). https://doi.org/10.1080/10357718.2021.1909534

European Parliament (2023). *MEPs ready to negotiate first-ever rules for safe and transparent AI.* European Parliament. https://acortar.link/a9Os2l

Ferraris, M. (2019). *Posverdad y otros enigmas*. Alianza.

Fontcuberta, J. (2017). *La furia de las imágenes. Notas sobre la postfotografía*. Galaxia Gutenberg.

Gamir-Ríos, J. y Tarullo, R. (2022). Predominio de las *cheapfakes* en redes sociales. Complejidad técnica y funciones textuales de la desinformación desmentida en Argentina durante 2020. *AdComunica*, 23, 97-118.

https://doi.org/10.6035/adcomunica.6299

Hameleers, M., Powell, T. E., Van Der Meer, T. G. L. A. y Bos, L. (2020). A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication*, *37*(2), 281-301. https://doi.org/10.1080/10584609.2019.1674979

Intel (2022). *Intel Introduces Real-Time Deepfake Detector*. https://acortar.link/6RKmBp

Krizhevsky, A., Sutskever, I. y Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. En F. Pereira, C.J.C. Burges, L. Botton y K.Q. Weinberger (Eds.), *Advances in neural information processing systems*, 25 (pp. 1097–1105). Currant Associates, Inc. https://acortar.link/jGDz7g

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P. y Schroeder, D. T. (2021). Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes. *Frontiers in Communication*, 6, 632317. https://doi.org/10.3389/fcomm.2021.632317

Lewandowsky, S., Cook, J. y Ecker, U. K. (2017). Letting the gorilla emerge from the mist: Getting past post-truth. *Journal of Applied Research in Memory and Cognition*, *6*(4), 418-424. https://doi.org/10.1016/j.jarmac.2017.11.002

Lilleker, D. G. y Liefbroer, M. (2018). 'Searching for something to believe in': Voter uncertainty in a post-truth environment. *International Journal of Media & Cultural Politics*, *14*(3), 351-366. https://doi.org/10.1386/macp.14.3.351_1

Lynch, M. P. (2016). *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data*. W. W. Norton & Company.

Marzal-Felici, J. (2021). Proposals for the study of images in the post-truth era. *El Profesional de la Información*, *30*(2). https://doi.org/10.3145/epi.2021.mar.01

McDermott, R. (2019). Psychological underpinnings of post-truth in political beliefs. *PS: Political Science & Politics*, *52*(2), 218-222. https://doi.org/10.1017/S104909651800207X

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A. y Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18. https://doi.org/10.1186/s12874-018-0611-x

Nelson, A. y Lewis, J. A. (2019). *Trust your eyes? Deepfakes policy brief*. Center for Strategic and International Studies.

Neuman, N., Fletcher, R., Eddy, K., Robertson, C. T. y Nielsen, R. K. (2023). *Reuters Institute Digital News Report 2023*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023

Neuman, N., Fletcher, R., Schulz, A., Andı, S. y Nielsen, R. K. (2020). *Reuters Institute Digital News Report 2020*. Reuters Institute for the Study of Journalism. https://acortar.link/p4gn2f

Pérez Dasilva, J., Meso Ayerdi, K. y Mendiguren Galdospin, T. (2021). Deepfakes on Twitter:

Which Actors Control Their Spread? *Media and Communication*, *9*(1), 301-312. https://doi.org/10.17645/mac.v9i1.3433

Prozorov, S. (2019). Why is there truth? Foucault in the age of post-truth politics. *Constellations*, 26, 18-30. https://doi.org/10.1111/1467-8675.12396

Qian, S., Shen, C. y Zhang, J. (2022). Fighting cheapfakes: Using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. *Journal of Computer-Mediated Communication*, *28*(1), zmac024. https://doi.org/10.1093/jcmc/zmac024

Shahid, F., Kamath, S., Sidotam, A., Jiang, V., Batino, A. y Vashistha, A. (2022). "It Matches My Worldview": Examining Perceptions and Attitudes Around Fake Videos. En *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-15) Association for Computing Machinery. https://doi.org/10.1145/3491102.3517646

Shin, S. Y. y Lee, J. (2022). The Effect of Deepfake Video on News Credibility and Corrective Influence of Cost-Based Knowledge about Deepfakes. *Digital Journalism*, *10*(3), 412-432. https://doi.org/10.1080/21670811.2022.2026797

Sontag, S. (1981). *Sobre la fotografía*. Edhasa.

Vaccari, C. y Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*(1). https://doi.org/10.1177/2056305120903408

Wardle, C. y Derakhshan, H. (2017). *Information Disorder. Toward an interdisciplinary framework for research and policy making (DGI [2017] 09)*. Council of Europe report. https://acortar.link/2sjVWZ

Weeks, Brian E. y Garrett, R. K. (2014). Electoral Consequences of Political Rumors: Motivated Reasoning, Candidate Rumors, and Vote Choice during the 2008 U.S. Presidential Election. En: *International Journal of Public Opinion Research*, *26*(4). Oxford University Press (pp. 401–422). https://doi.org/10.1093/IJPOR/EDU005

Yadlin-Segal, A. y Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence*, *27*(1), 36-51. https://doi.org/10.1177/1354856520923963

CONTRIBUCIONES DE AUTORES, FINANCIACIÓN Y AGRADECIMIENTOS

Conceptualización: Ballesteros Aguayo, Lucía y Ruiz del Olmo, Francisco Javier. **Metodología:** Ruiz del Olmo, Francisco Javier y Ballesteros Aguayo, Lucía. **Validación**: Ruiz del Olmo, Francisco Javier y Ballesteros Aguayo, Lucía. **Análisis formal**: Ballesteros Aguayo, Lucía y Ruiz del Olmo, Francisco Javier. **Curación de datos**: Ballesteros Aguayo, Lucía y Ruiz del Olmo, Francisco Javier. **Redacción-Preparación del borrador original**: Ballesteros Aguayo, Lucía y Ruiz del Olmo, Francisco Javier. **Redacción-Revisión y Edición**: Ballesteros Aguayo, Lucía y Ruiz del Olmo, Francisco Javier. **Visualización**: Ballesteros Aguayo, Lucía y Ruiz del Olmo, Francisco Javier. **Supervisión:** Ballesteros

Aguayo, Lucía. **Administración de proyectos**: Ballesteros Aguayo, Lucía. **Todos los autores han leído y aceptado la versión publicada del manuscrito:** Ballesteros Aguayo, Lucía y Ruiz del Olmo, Francisco Javier.

Financiación: El presente texto nace en el marco del proyecto Posverdad a debate: reconstrucción social tras la pandemia (P020_00703).

Agradecimientos: A los miembros del Proyecto Posverdad a debate: reconstrucción social tras la pandemia (P020_00703).

Lucia Ballesteros Aguayo
Universidad de Málaga.
Doctora Sobresaliente *Cum laude* y Mención Internacional, fruto de la obtención de Ayudas Internacionales para la realización de estancias de investigación en la Universidad La Sapienza (Roma) y en la Universidad de Aveiro (Portugal). Acreditada por la ANECA como Contratado Doctor es Licenciada en Periodismo y en Publicidad y RRPP y Máster en Formación del Profesorado por la Universidad Complutense de Madrid. Actualmente es profesora en la Facultad de Comunicación de la Universidad de Málaga y Coordinadora Docente de asignaturas basadas en las TIC. Líneas de investigación: Communication, Disinformation, Propaganda y Media Studies.
Índice H: 4.
Google Scholar: https://scholar.google.com/citations?user=wcxyZPUAAAAJ&hl=en
Orcid ID: https://orcid.org/0000-0003-1191-4070
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57211473463

Francisco Javier Ruiz del Olmo
Universidad de Málaga.
Catedrático de la Universidad de Málaga, en el área de Comunicación Audiovisual; su labor docente e investigadora se desarrolla en las Facultades de Ciencias de la Comunicación y en Bellas Artes. Ha investigado los modelos comunicacionales y las formas contemporáneas de los medios, así como los usos técnicos y sociales de los mismos; una segunda línea de investigación que desarrolla está relacionada con la comunicación y los nuevos medios. En la actualidad codirige el Proyecto de Investigación de Generación de Conocimiento del Ministerio de Ciencia e Innovación "AV-ADJUVEN: Repertorios y prácticas mediáticas en la adolescencia y la juventud: recepción y uso de plataformas audiovisuales online" (PID2022-138281NB-C22).
Índice H: 12.
Google Scholar: https://scholar.google.es/citations?user=RphOTRkAAAAJ&hl=es&oi=ao
Orcid ID: https://orcid.org/0000-0002-1953-1798
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=55581055500

Revista de Ciencias de la Comunicación e Información
ISSN: 2695-5016

ARTÍCULOS RELACIONADOS:

Anton-Bravo, A. y Serrano Tellería, A. (2021). Innovación en la docencia del periodismo a través de la ciencia de datos. *European Public & Social Innovation Review*, *6*(1), 70-84. https://pub.sinnergiak.org/esir/article/view/150

Encinillas García, M. y Martín Sabarís, R. (2023). Desinformación y Salud en la era PRECOVID: Una revisión sistemática. *Revista de Comunicación y Salud*, 13, 1-15. https://doi.org/10.35669/rcys.2023.13.e312

Fernández Fernández, M. (2021). Informaciones, crónicas y relaciones del siglo XVI como antecedentes de las fake news. El caso de la "invención" de San Segundo. *Historia y Comunicación Social*, *26*(2), 593-601. https://doi.org/10.5209/hics.68530

Toro González, S. y Pérez-Curiel, C. (2021). Populismo político en tiempos de COVID. Análisis de la estrategia de comunicación de Donald Trump y Boris Johnson en Twitter. *Revista de Comunicación de la SEECI*, 54, 1-24. https://doi.org/10.15198/seeci.2021.54.e700

Zúñiga, F., Mora Poveda, D. A. y Molina Mora, D. P. (2023). La importancia de la inteligencia artificial en las comunicaciones en los procesos marketing. *Vivat Academia*, 156, 19-39. https://doi.org/10.15178/va.2023.156.e1474