Artificial intelligence, disinformation and media literacy proposals around deepfakes

Inteligencia artificial, desinformación y propuestas de alfabetización en torno a los deepfakes

Miriam Garriga*, Raquel Ruiz-Incertis **, Raúl Magallón-Rosa**

* Communication Department, University Carlos III Madrid (mgarriga@pa.uc3m.es)
** Communication Department, University Carlos III Madrid (raquel.r.incertis@alumnos.uc3m.es)
*** Communication Department, University Carlos III Madrid (raul.magallon@uc3m.es)

Abstract

The role of artificial intelligence and its place in the new disinformation strategies is perhaps one of the most difficult issues to focus on nowadays, since we are at the beginning of a process of definition and ways of exploration. In this paper, first of all, we analyze the different approaches that are being applied to the regulation of artificial intelligence and that may affect the different disinformation strategies that are being identified. Secondly, we study how artificial intelligence is being used to identify disinformation content. In this regard, from the point of view of verification processes, one of the main challenges is when identifying deepfakes (images and video, mainly) linked to news cycles. From this perspective, a typology of deepfakes is proposed and its main characteristics will be described according to the verifications carried out by the Spanish fact-checking organizations. Finally, a set of recommendations will be presented to work from a media literacy point of view with the identification of deepfakes.

Keywords: deepfakes, disinformation, fact-checking, Spain, media literacy

Resumen

El papel de la inteligencia artificial y su lugar en las nuevas estrategias de desinformación es quizá una de las problemáticas más complicadas de abordar en la actualidad, ya que estamos en el principio de un proceso que establecerá ramificaciones, limitaciones y tendrá que solucionar problemas adquiridos previamente por otras tecnologías a medida que se vaya desarrollando. En primer lugar, analizamos los distintos enfoques que se están aplicando en torno a la regulación de la inteligencia artificial y que pueden afectar a las distintas estrategias de desinformación que se van identificando. En segundo lugar, estudiamos cómo se está utilizando la inteligencia artificial para identificar contenidos desinformativos. Al respecto, es obvio que, desde el punto de vista de la verificación, uno de los principales desafíos se establece a la hora de identificar *deepfakes* (imágenes y vídeos, principalmente) vinculados a ciclos de actualidad. Desde esta perspectiva, se propondrá una tipología de *deepfakes* y se describirán sus características principales atendiendo a las verificaciones realizadas por las organizaciones de fact-checking españolas. Por último, se presentarán una serie de recomendaciones para trabajar desde la alfabetización mediática en la identificación de *deepfakes*.

Palabras clave: deepfakes, desinformación, fact-checking, España, alfabetización mediática

Introducción

El papel de la inteligencia artificial y su aplicación a la lucha contra la desinformación es una de las problemáticas más complicadas de abordar en la actualidad, puesto que estamos en el principio de un proceso disruptivo y tecnológico que establecerá ramificaciones y limitaciones legislativas, sociales y educativas a medida que se vaya desarrollando (Hrckova *et al.*, 2022; DSN, 2023).

En este trabajo, analizamos de forma resumida los distintos enfoques que se están aplicando en torno a la regulación y adaptación de la inteligencia artificial y que pueden afectar a las distintas estrategias de desinformación que se van identificando. Para la UE el término "inteligencia artificial" (IA) se aplica a los sistemas que manifiestan un comportamiento inteligente, pues son capaces de analizar su entorno y pasar a la acción - con cierto grado de autonomía - con el fin de alcanzar objetivos específicos (Comisión Europea, 2018).

Posteriormente, en nuestra investigación identificamos cómo se está trabajando desde las organizaciones de fact-checking con herramientas de inteligencia artificial y los principales usos periodísticos que se han desarrollado en estos últimos años (Adair, 2021; Moran Shaikh, 2022 y Munoriyarwa et alt., 2023).

Un buen ejemplo son los discursos de representantes políticos, su transcripción y la incorporación y categorización de sus afirmaciones a distintas bases de datos para confirmar si se tratan de declaraciones verificables y, posteriormente, reconocer cuáles son las fuentes necesarias para verificar las afirmaciones categorizadas (Adair, 2020).

En este sentido, destacan también proyectos vinculados al aprendizaje automático que sirven para desarrollar mecanismos de alerta temprana ante acontecimientos con una fuerte carga de incertidumbre, ruido y miedo como son las catástrofes naturales o los atentados (Shin y Chan-Olmsted, 2022).

Por otra parte, se estudian iniciativas que están ayudando a agilizar el trabajo de las organizaciones de fact-checking como chatbots, programas de categorización de declaraciones, etc. (Babakar y Moy, 2016, Lim y Perrault, 2023). Desde esta perspectiva, hay que recordar que el fact-checking automatizado se concentra principalmente en tres acciones: identificación -a través de la monitorización-, verificación y corrección -con falsedades repetidas, proveyendo datos contextuales, etc.- (Graves, 2018).

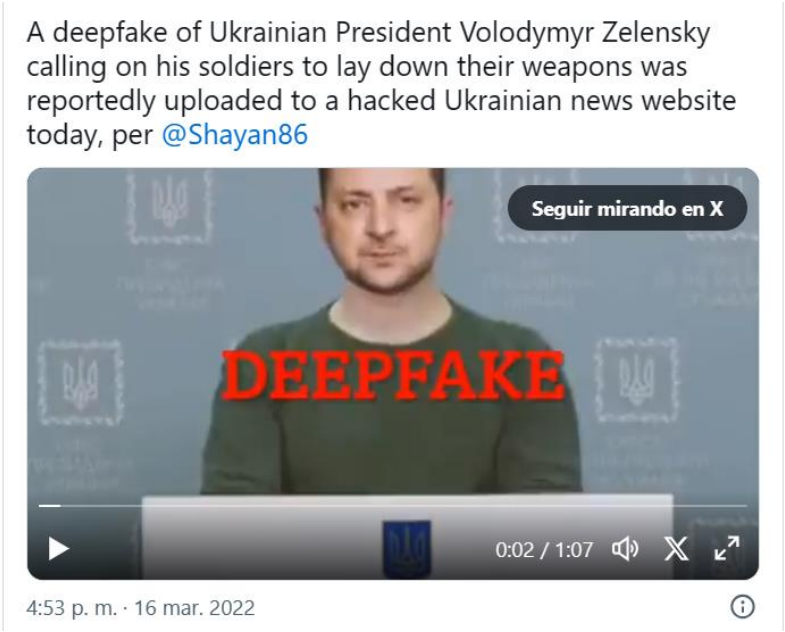
En tercer lugar, estudiamos cómo se está utilizando la inteligencia artificial para identificar contenidos desinformativos. Al respecto, es obvio que -desde el punto de vista de la verificación- uno de los principales desafíos se establece a la hora de identificar *deepfakes* e imágenes vinculadas a ciclos de actualidad informativos (Boté-Vericad y Vállez, 2022).

Se entienden por *deepfakes* aquellos archivos de vídeo, imagen o voz manipulados mediante una herramienta o programa dotado de tecnología de inteligencia artificial que permite el intercambio de rostros en imágenes y la modificación de la voz, de modo que los archivos parezcan originales, auténticos y reales. Para Fundéu, que prefiere hablar de "ultrafalso", la palabra *deepfake* "alude a los sistemas informáticos que permiten, mediante técnicas de inteligencia artificial, desarrollar vídeos manipulados extremadamente realistas, aunque también es frecuente que se aplique a los vídeos así creados. El realismo es tal que puede ser imposible saber que ha sido falseado, lo que sirve, por ejemplo, para propagar contenidos falsos con apariencia de noticias y como pornovenganza. Este es su uso original, pero en ocasiones se utiliza para manipulaciones similares, como en audio".[1] En este trabajo utilizaremos el término *deepfake* por su uso más

[1] Véase: https://www.fundeu.es/recomendacion/ultrafalso-alternativa-a-deep-fake/

extendido y por servir de paraguas conceptual para analizar la relación entre inteligencia artificial y el uso de vídeos y audios de manera instrumentalizada a través de su manipulación.

Imagen 1. Ejemplo de *deepfake* del presidente de Ucrania al inicio de la invasión rusa en marzo de 2022.



Fuente: Twitter/Maldita.

Los *deepfakes* se utilizan para inducir a engaño a las personas receptoras, ya sea haciendo que un representante político diga algo en un vídeo que realmente nunca dijo con el propósito de interferir en una campaña política o incluyendo la imagen de un famoso (o de cualquier persona) en un material pornográfico con el objetivo de perjudicarle o chantajearle (Cerdán y Padilla, 2019), por lo que suponen una gran amenaza desinformativa para la sociedad actual (INCIBE, 2022).

Revisión de literatura. La incipiente regulación de la Inteligencia artificial en la UE y EE. UU

La regulación de la inteligencia artificial se ha convertido en una cuestión geopolítica que puede definir económica, tecnológica y culturalmente avances y retrocesos sociales de la próxima década. Las medidas que se tomen -como ya ocurrió con la universalización progresiva de internet o la proliferación de las redes sociales- pueden tener consecuencias de carácter democrático, socioeducativo y económico.

En el caso de las redes sociales, sólo con la *Digital Service Act* aprobada en 2022 por parte de la UE se empezó realmente a poner en cuestión el papel de las plataformas como editores que pueden determinar el alcance de los contenidos. En diciembre de 2023, se anunció que el Consejo y el Parlamento Europeo habían llegado a un acuerdo provisional sobre la propuesta que establece normas armonizadas sobre inteligencia artificial (IA), también llamada *Ley de inteligencia artificial*.

En el caso de EEUU, recientemente se intentó que el Tribunal Supremo cambiara de posición en torno al artículo 230 de la *Ley de Decencia en las Comunicaciones de 1996*.

Un artículo en el que tradicionalmente se han amparado las redes sociales y empresas tecnológicas para que no fueran consideradas responsables de los contenidos que circulan en sus plataformas y que establece:

"Ningún proveedor o usuario de un servicio informático interactivo será tratado como editor o difusor de información".

En este sentido, hemos de señalar que la problemática que engloba al desarrollo de la inteligencia artificial se ha visto cómo una cuestión que necesitaba ser regulada desde casi el principio. Sin embargo, el enfoque de cómo afrontar su uso para campañas de desinformación está determinado por cómo puede afectar económica y tecnológicamente su adaptación al uso cotidiano de la sociedad.

Al respecto, el año 2016 fue fundamental a la hora de establecer las bases de reflexión regulatoria de la inteligencia artificial. La *Oficina de Política Científica y Tecnológica* de la Casa Blanca, la *Comisión de Asuntos Jurídicos del Parlamento Europeo* y, en el Reino Unido, el *Comité de Ciencia y Tecnología de la Cámara de los Comunes* publicaron sus informes iniciales sobre cómo prepararse para el futuro de la IA.

Ya en 2017, la UE dio un primer paso para proponer el establecimiento de una personalidad electrónica que sirviera como base para un régimen jurídico de responsabilidad en caso de perjuicios causados por acciones imprevisibles de sistemas de inteligencia artificial que fueran capaces de tomar decisiones de manera independiente, inteligente o que se comunicaran con terceros de forma autónoma (Parlamento Europeo, 2017).

Por su parte, la Comisión Europea para la Eficiencia de la Justicia (CEPEJ) elaboró en 2018 la *Carta Europea de Ética para el uso de la Inteligencia Artificial en los Sistemas Judiciales y su Entorno*. Sin embargo, no es hasta el año 2020 cuando el Parlamento Europeo advierte claramente de los peligros que puede entrañar la inteligencia artificial aplicada a fines desinformativos. Concretamente, en una Resolución sobre aspectos éticos de esta tecnología, este organismo afirma que Europa podría verse expuesta a:

> *La explotación de sesgos en los datos y los algoritmos por terceros países a través de burbujas informativas, con el objetivo de crear perfiles manipulables de forma malintencionada; facilitando así la distribución de noticias falsas*.

Para luchar contra este prejuicio, la UE propone:

> *Seguir invirtiendo en investigación, con el fin de desarrollar tecnologías de IA que combatan la desinformación en consonancia con el Derecho de la Unión* (algoritmos con transparencia y rendición de cuentas) *a fin de garantizar el acceso a contenidos diversos desde el punto de vista cultural y lingüístico.*

Además, se subraya específicamente que "*las IAs no podrán interferir en elecciones ni contribuir a la difusión de desinformación"* (Parlamento Europeo, 2020).

En la actualidad, la Unión Europea busca liderar la regulación de la inteligencia artificial (IA). Ésta es considerada una innovación tecnológica que los 27 consideran tanto una oportunidad como un desafío en términos de seguridad. Dos años de intensas negociaciones, dieron forma a una normativa cuya meta es prevenir situaciones similares a las ocurridas con el desarrollo de Internet y el mal uso de los datos personales (Del Castillo y Castro, 2023).

Esta futura ley pionera establece, entre otras cuestiones, una escala de riesgos con respecto al uso de la IA, desde los menos peligrosos (restringidos) hasta los que la legislación europea califica como "inasumibles", los cuales quedarán prohibidos en territorio comunitario (Mortera-Franco, 2023). La propuesta de ley de la IA incluye, sin embargo, algunas excepciones.

Mientras que el Parlamento Europeo abogaba por una prohibición total de estos sistemas de hipervigilancia, los Estados miembros exigieron introducir algunas cláusulas vinculadas a la seguridad nacional.

En cuanto a Estados Unidos, este país ha publicado un menor número de leyes referentes a la inteligencia artificial que la Unión Europea. Además, la legislación estadounidense que se creó hasta el año 2021 estaba fragmentada y principalmente derivaba de las leyes y normativas aprobadas por cada Estado de forma individual. Estas regulaciones se centraron en la creación de comités para decidir la forma en la que las instituciones públicas podían utilizar la IA e investigar sus consecuencias[2], siendo Nueva York el estado que más legislación publicó sobre este asunto[3].

Con el fin de reforzar y coordinar las actividades de investigación, desarrollo, demostración y enseñanza de la IA en todos los Departamentos y Agencias de EE.UU., en 2022 se presentó tanto al Senado como a la Cámara de Representantes la *Ley de Responsabilidad Algorítmica*[4], la cual tenía la intención de instruir a la *Agencia Nacional de Protección del Consumidor* (CFC) para que desarrollara normativas que exigieran a las empresas analizar las consecuencias de la comercialización de sus tecnologías de IA.

En este contexto, recientemente se ha producido un cambio de rumbo en el que existe una mayor conciencia sobre los peligros que puede suponer la IA, estableciéndose así el *AI Bill of Rights*[5], sucesora del *Artificial Intelligence for the American People*[6].

En resumen, mientras que las políticas legislativas de la UE están más orientadas a la protección del usuario, en EEUU la perspectiva hacia esta tecnología y sus usos se traduce también en aprovechar los beneficios comerciales y geopolíticos de la IA.

En Europa, la utilización de la inteligencia artificial conllevará una serie de controles más estrictos, así como un análisis continuado por parte tanto de las autoridades nacionales como de la nueva oficina de inteligencia artificial europea.

La nueva ley de regulación de la IA, no obstante, pretende evitar entorpecer el desarrollo de esta tecnología, por lo que intentará combinar la protección de los derechos del ciudadano con el progreso técnico.

Asimismo, buscará proteger ciertos derechos sectoriales de sustancial importancia, como es el caso de los derechos de autor (Mortera-Franco, 2023).

Periodismo, fact-checking e inteligencia artificial. Usos y funcionalidades incipientes

---

[2] National Conference of States Legislatures (2022). *Legislation Related to Artificial Intelligence.* https://www.ncsl.org/technology-and-communication/legislation-related-to-artificial-intelligence

[3] U.S. Chamber of Commerce (2022). *State-by-State Artificial Intelligence Legislation Tracker. Interactive map shows states' action to legislate artificial intelligence.* https://www.uschamber.com/technology/state-by-state-artificial-intelligence-legislation-tracker?state=

[4] https://www.congress.gov/bill/117th-congress/house-bill/6580/text

[5] Office of Science and Technology Policy (2022). *Blueprint for an AI Bill of Rights. Making automated systems work for the american people.* The White House https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[6] The White House (2021). *Artificial Intelligence for the American People*. https://trumpwhitehouse.archives.gov/ai/

El uso de la inteligencia artificial y la automatización en el periodismo lleva años extendiéndose y, cada vez más medios de comunicación, están utilizando este tipo de tecnología para producir noticias de manera más eficiente. Al respecto, es creciente el número de las organizaciones periodísticas que utilizan la IA con el objetivo de generar material para las noticias utilizando datos y algoritmos (De-Lima-Santos, & Salaverría, 2021; De-Lima-Santos & Cerón, 2021).

Desde esta perspectiva, las herramientas de inteligencia artificial se están aplicando desde hace algún tiempo en el periodismo en tareas como la transcripción, la traducción, la revisión gramatical, la clasificación de contenido, el reconocimiento de imágenes, la personalización de contenido, el intento de optimización de ingresos, etc. (Ulken, 2022).

En la misma línea, autores como Pellicer destacan que los usos vinculados a la creación periodística son claros: creación de contenidos, desarrollo de textos vinculados a redes sociales, potenciación de artículos a través de imágenes (ilustraciones, fotos o vídeos) que usan en su creación la IA; desarrollo de chatbots de atención al cliente, automatización de procesos periodísticos, verificación de contenidos, etc. (Pellicer, 2022).

De esta manera, también está siendo transformada la forma en que se produce y se consumen contenidos procedentes de los medios de comunicación. Por lo tanto, la capacidad de procesar grandes cantidades de datos, identificar patrones y realizar predicciones son algunas de las ventajas de la IA en el campo periodístico.

En el campo de la desinformación los peligros que suponen las IAs están por definir, pero también por delimitar (Funke, 2018; Araujo et alt., 2020). Con la capacidad de producir y difundir contenido de forma masiva, determinados actores generadores de contenidos falsos pueden usar esta tecnología para crear narrativas engañosas que se difundan rápidamente y alcancen a audiencias masivas.

Además, los algoritmos de IA pueden ser manipulados para impulsar la difusión de contenidos falsos a través de la personalización de contenido y la orientación de audiencias específicas.

En este contexto, cada vez son más frecuentes proyectos que pretenden desarrollar herramientas que permitan a los periodistas, los responsables políticos y al público en general distinguir entre información fiable y contenidos falsos. Para ello, se están utilizando técnicas de procesamiento del lenguaje natural (PLN) y minería de datos con el objetivo de analizar grandes cantidades de información y determinar su veracidad.

Al respecto, las últimas iniciativas dirigidas a combatir la desinformación, respaldadas por las instituciones públicas tanto europeas como nacionales, coinciden en la inclusión de la alfabetización mediática como un elemento clave (Sánchez-Illán, 2021; DSN, 2023). De hecho, algunas propuestas teóricas recientes han apuntado hacia la idea de expandir este concepto hacia lo que se denomina "alfabetismo transmedia". Esto implica considerar al consumidor de información como un "actor activo que no solo adquiere habilidades cada vez más avanzadas para entender los nuevos formatos narrativos, sino que también contribuye cada vez más a la creación, combinación y compartición de contenido en las redes digitales" (Sádaba y Salaverría, 2023, p. 21).

En el caso de organizaciones de fact-checking como Chequeado, la mitad de las verificaciones de líderes políticos que hicieron en 2020, nacieron de frases encontradas gracias a Chequeabot, la plataforma de automatización que desarrollaron en base a tecnologías de procesamiento del lenguaje natural y *machine learning* (Fernández, 2021).

Previamente, en 2018 hicieron una verificación en vivo del debate en el Senado sobre la legalización de la interrupción voluntaria del embarazo y pudieron por primera vez tener una transcripción en tiempo real de

todo lo que se dijo en la sesión. El desafío de tomar distintas voces en un contexto diferente como el de una sesión del Senado lo testearon la semana anterior al debate y ahí pudieron comprobar que también funcionaba correctamente (Fernández, 2018).

Otro ejemplo es el de ClaimBuster, una plataforma desarrollada en la Universidad de Texas-Arlington y que fue entrenada con hasta 20.000 frases de pasados debates presidenciales de EE. UU. (Hassan et al. 2017).

En España, uno de los ejemplos más destacados es el de ClaimCheck de Newtral. Una herramienta (*ClaimHunter*) detecta automáticamente las afirmaciones políticas realizadas en Twitter de determinadas cuentas seleccionadas, mientras que otra aplicación transcribe a texto la cobertura de video y audio de los representantes políticos. De este modo, las dos aplicaciones *dialogan* e "identifican y resaltan declaraciones que contienen una afirmación relevante para la vida pública que se puede probar o refutar, como en declaraciones que no son ambiguas (...) y se las envían a los verificadores de hechos de *Newtral* para su revisión" (Morrish, 2023).

El sistema, y es importante subrayar que no es sencillo, pretende distinguir opiniones de hechos. Como recuerda Marilín Gonzalo: "en Newtral llevamos desde 2019 desarrollando herramientas para la monitorización y análisis del discurso público en redes sociales. Con Claimhunter detectamos frases verificables en Twitter dichas por políticos y con ClaimCheck analizamos si una frase dicha por ellos ya ha sido repetida por otros en el pasado" (Gonzalo, 2023).

En este recorrido sobre la evolución de la IA, una de las consecuencias de la pandemia fue que -tras el aumento exponencial de desinformación- las organizaciones de fact-checking vieron claramente la necesidad de automatizar lo máximo posible tanto los procesos de identificación de bulos repetidos como la capacidad de respuesta en plataformas como WhatsApp. En el caso de organizaciones como Maldita, con la llegada de la pandemia de COVID-19 el volumen de consultas se multiplicó y pasaron de 200-300 consultas diarias a 2.000. Es decir, casi se multiplicó por 10.

En este escenario, y desde junio de 2020, Maldita.es cuenta con un servicio automatizado de verificación en WhatsApp que permite comparar contenidos enviados por los usuarios con su base de datos y responder de manera automática si se confirma que existe un desmentido[7]. En su primer año de funcionamiento más de 26.000 personas lo utilizaron, de las cuales el 61% lo hizo más de una vez. La automatización permitió enviar más de 400.000 mensajes y verificar 108.000 contenidos (Maldita, 2021).

Al respecto, hay que señalar que esta automatización de determinadas fases del proceso de verificación ha permitido a las organizaciones de fact-checking mejorar tanto el tiempo de verificación como la curva de identificación de posibles mensajes verificables, cruzar grandes bases de datos y automatizar determinados procesos relacionados con la genealogía de la desinformación, complementar la información con gráficos, etc.

Metodología y preguntas de investigación. ¿Cómo analizar la verificación de los *deepfakes*?

[7] Véase: https://api.whatsapp.com/send/?phone=34644229319&text&type=phone_number&app_absent=0

Como señalamos en la introducción, uno de los objetivos de este trabajo es analizar las problemáticas que abre un fenómeno de interés creciente -como son los *deepfakes*- al trabajo de las organizaciones de fact-checking. Para ello, se estudian las piezas relacionadas con los *deepfakes* así como los desmentidos realizados por las principales organizaciones de fact-checking españolas hasta las elecciones autonómicas y municipales del 28 de mayo de 2023. La elección de la fecha estaba determinada por el posible uso novedoso que podían tener este tipo de contenidos en los procesos electorales, pero también por el aumento de recursos que las organizaciones de fact-checking aportan durante las campañas electorales (Rocha et alt., 2019).

Para el análisis de contenido, se seleccionaron dos unidades de análisis: los artículos redactados por los *verificadores* y los bulos unívocos -no repetidos- identificados y publicados entre febrero de 2021 y mayo de 2023. En este sentido, se buscaba también identificar cuántas piezas informativas eran explicativas y cuántas estaban relacionadas con verificaciones concretas.

Desde un punto de vista temporal se localizaron 3 piezas en el año 2021, 11 en el año 2022 y 21 en el año 2023, por lo que se observa un aumento progresivo de este tipo de formatos.

Al respecto, las organizaciones de fact-checking analizadas fueron Maldita.es, EFE Verifica, Verifica RTVE, Newtral y Verificat.cat, todas ellas con una consolidación creciente en los últimos años. Desde esta perspectiva, hay que destacar que son miembros de la International Fact-checking Network (IFCN) y cumplen con sus criterios de calidad, EFE Verifica, Newtral, Verificat y Maldita.[8]

En el segundo caso, el de los bulos unívocos, se estudiaron las verificaciones de *deepfakes* individuales. Es decir, que no se repiten entre las distintas organizaciones de fact-checking ni tampoco se agrupan varios en una única pieza informativa.

Imagen 2. Ejemplo de *deepfake* unívoco que no se repite entre las organizaciones de fact-checking españolas

A la izquierda, fotograma del vídeo manipulado con los datos alterados. A la derecha, la emisión original de la cadena TV Globo | Fuente: Agencia LUPA

Fuente: Newtral/Agencia Lupa.

[8] Véase: https://ifcncodeofprinciples.poynter.org/

En este contexto, y para analizar el fenómeno de los *deepfakes* y sus formas de verificación, las preguntas de investigación que se plantearon fueron:

1. ¿Cuántos *deepfakes* están relacionados con cuestiones de actualidad internacional y cuántos son de ámbito local?

2. ¿Cuáles son los formatos más utilizados?

3. ¿Qué tipo de género periodístico/informativo utilizan las organizaciones de fact-checking para la publicación de las verificaciones sobre *deepfakes*? ¿Hay alguna diferencia?

4. ¿Qué características comunes tienen los protagonistas de las verificaciones?

5. ¿Cuáles son las temáticas preferentes?

6. ¿Qué características comunes explican la viralidad de los *deepfakes*?

7. ¿Hay diferencias entre los deepfakes relacionados con hombres y con mujeres?

Para el análisis, inicialmente se hizo una distinción entre los distintos formatos: imagen, vídeo y audio. Posteriormente se realizó una categorización por temáticas, distinguiendo entre: religión, guerra rusoucraniana, política, consumo de drogas entre personalidades políticas, pornografía y otros.

También se realizó un análisis sobre las principales características de viralidad identificadas: políticos involucrados, celebridades implicadas (incluye figuras religiosas), temática pornográfica, eventos de actualidad y otros (como por ejemplo cuestiones de carácter estético).

Por último, se estudiaron los principales elementos de verificación identificados. Entre ellos destacaban las alteraciones en la imagen propias de las IA (varios dedos, orejas grandes…), faciales y de iluminación, ralentización de movimientos, incoherencias en la narrativa propia de un personaje conocido y otros (experiencias similares, plantilla de aplicaciones, postproducción, etc.).

Análisis de los resultados. Tipologías y desafíos que presentan los *deepfakes* a la verificación.

Como hemos señalado en el apartado metodológico, para la realización del trabajo de investigación sobre el contenido relacionado con los *ultrafalsos* o *deepfakes* se seleccionaron dos unidades de análisis: los artículos redactados por los *verificadores* y los bulos unívocos identificados por el conjunto de las organizaciones periodísticas entre febrero de 2021 y mayo de 2023, hasta las elecciones autonómicas y municipales españolas del 28 de mayo de 2023.

En el primer caso, la muestra es de 35 publicaciones -incluidas agrupaciones de bulos sobre *deepfakes*-; y en el segundo, de 33 verificaciones. En referencia a los artículos redactados, se encontraron 35 piezas informativas que pueden dividirse en explicativas y de verificación. Las piezas explicativas hacen referencia a la definición, características, etc. de los *deepfakes* y las de verificación a piezas concretas vinculadas a alguna imagen, vídeo o audio creado con inteligencia artificial. Del total de 35 piezas identificadas, 10 fueron explicativas y 25 estaban relacionadas con verificaciones concretas.

Por otra parte, y en base a los resultados obtenidos, se puede determinar que Newtral es la organización periodística que más ha centrado su atención en los bulos generados a través de inteligencia artificial o *deepfakes* (Gráfico 1). De hecho, casi la mitad han sido desmentidos por esta organización (48,6%). Le sigue Maldita.es, con un 34,3%; y en un porcentaje más reducido EFE Verifica (11,4%).

Es reseñable que tanto Verifica RTVE como Verificat, hasta mayo de 2023 sólo habían publicado un artículo sobre este asunto. Además, en ambos casos, había sido más con un objetivo explicativo -relatar a sus lectores cuáles eran las últimas tendencias en creación de bulos-, que verificativo -exponer las verificaciones a través de diferentes fuentes-. En cualquier caso, estas tendencias demuestran que cada verificador posee la potestad de desarrollar su propia agenda mediática, estableciendo cuáles son aquellas cuestiones que desea introducir entre sus publicaciones.

Gráfico 1. Artículos de verificación relacionados con los *deepfakes* publicados por cada organización de fact-checking.

Fuente: Elaboración propia.

Destaca también el tiempo de verificación de estos *deepfakes*. En el 80% de las informaciones no se identifica el tiempo de verificación de manera precisa y en el 20 por ciento restante se identifica en las primeras 24 horas.

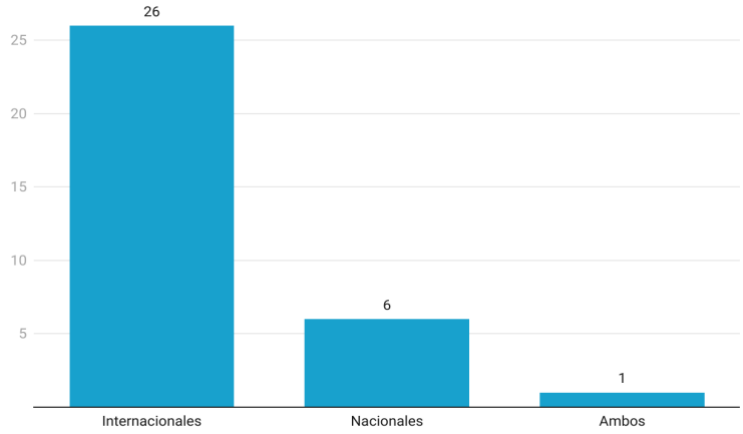
Características de los bulos analizados por los fact-checkers

En este apartado se analizan las verificaciones de *deepfakes* individuales unívocos, es decir, que no se repiten entre las distintas organizaciones de fact-checking ni tampoco se agrupan varios en una única pieza informativa. El objetivo es estudiar el tipo de formato, las temáticas, las particularidades del sujeto afectado por el bulo, etc. Es por ello por lo que la muestra ve alterado su número, que pasa de 35 a 33 publicaciones.

Al respecto, cabe destacar que casi un tercio de los artículos publicados (28,6%) son agrupaciones de bulos.

Por otra parte, y en la mayoría de los casos (Gráfico 2), los *deepfakes* son de carácter internacional (78,8%). Es decir, las historias que narran no se desarrollan en España, ni sus protagonistas tienen dicha nacionalidad. Por tanto, es probable que dichos bulos hayan sido de algún modo "exportados", y que hayan llegado a España a través de traducciones, ya sea porque en ellos aparecen personajes conocidos o porque traten eventos o noticias populares o mediáticas internacionalmente. En cuanto a los sujetos de los bulos nacionales, cabe mencionar que de los seis analizados cinco de ellos corresponden a representantes políticos y uno pertenece a la monarquía.

Gráfico 2. Número de bulos analizados según su nacionalidad.

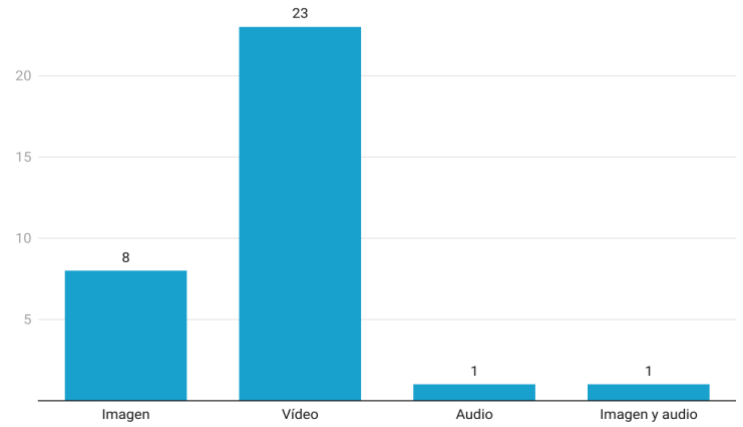


Fuente: Elaboración propia

En el *gráfico 3*, pueden observarse los formatos más utilizados por los creadores de *deepfakes* durante el periodo analizado. El vídeo es el formato preferido (66,6%), seguido de la imagen (24,2%). El audio tiene una aparición esporádica, dado que solo aparece en una ocasión.

Igualmente, se ha querido dejar evidencia de que existen ciertos bulos desarrollados por inteligencia artificial que combinan dos formatos. Es el caso del *deepfake* en el que se difunde una pieza informativa de telediario manipulada, en la que se modifica una imagen -correspondiente a un gráfico de intención de voto- y una voz en off -correspondiente a la presentadora que narra la noticia-. En este caso, el afectado es el entonces candidato a la presidencia brasileña Lula da Silva, que era el favorito en las encuestas y cuyo porcentaje fue reemplazado por el de su opositor Jair Bolsonaro[9].

Gráfico 3. Número de bulos analizados según su formato



Fuente: Elaboración propia

[9] Infantes Capdevila, G. (2022). Un 'deepfake' altera un informativo brasileño para situar a Bolsonaro como favorito en las encuestas. *Newtral*. https://www.newtral.es/deepfake-bolsonaro-renata-vasconcellos/20220819/

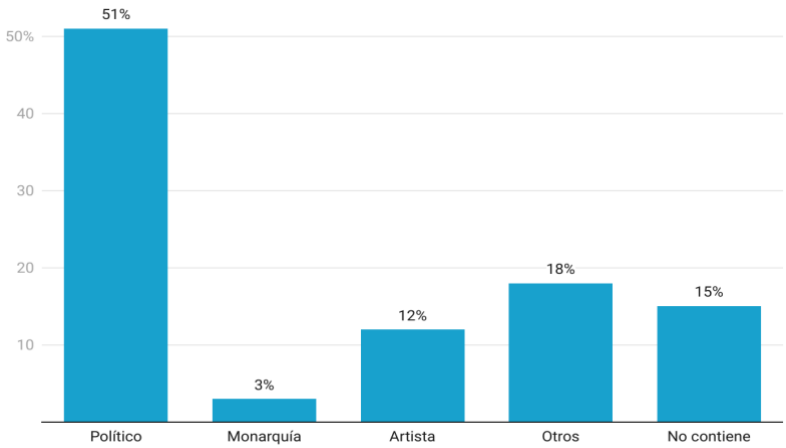
En el gráfico 4 puede observarse el porcentaje de bulos analizados según la profesión del sujeto protagonista del *deepfake* en base a cinco categorías. Sin embargo, con el objetivo de delimitarlas, se examinó previamente quiénes eran exactamente los individuos que constituían cada bulo. De las 36 personalidades categorizadas, 23 hacían referencia a hombres y 13 a mujeres.

Por otra parte, hay que subrayar que son las personalidades vinculadas a la política las más recurrentes. Hasta el 58,3% frente a artistas (19,4%) u *Otros,* donde se integran personalidades como Elon Musk, Bill Gates, Greta Thunberg o el Papa Francisco (16,6%) y un 5,5% vinculado a personalidades relacionadas con las distintas monarquías.

Ahora bien, con el objetivo de centrarnos en la unidad de análisis especificada, que son los 33 *deepfakes* analizados en el periodo previamente mencionado, se decidió incorporar los datos que se muestran en el Gráfico 4, los cuales están medidos sobre el total de bulos unívocos analizados y no sobre el total de las personalidades analizadas.

Si por ejemplo un bulo trata sobre Vladimir Putin y Xi Jingping, tan solo se contabilizará una vez la profesión "político". No obstante, tal y como se puede observar, los datos no muestran resultados muy diferentes.

Gráfico 4. Porcentaje de bulos analizados según la profesión del sujeto protagonista



Fuente: Elaboración propia

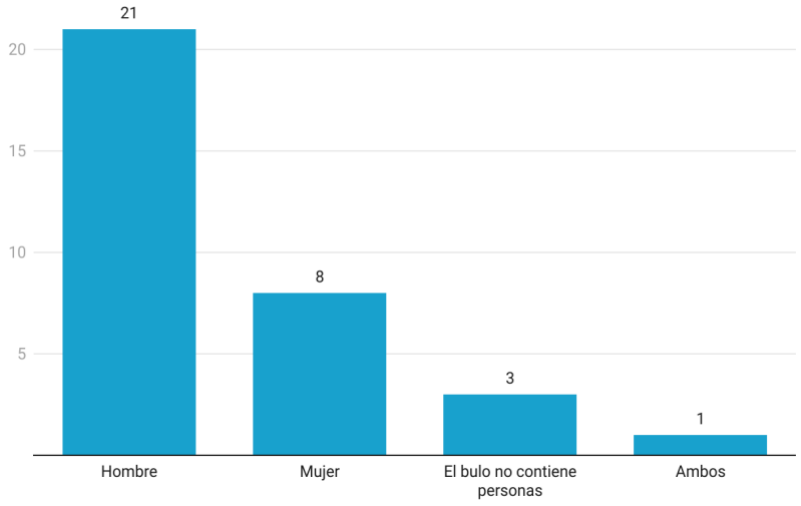
En cuanto al género (Gráfico 5), sí se puede observar una gran representación masculina (63,3%) frente a la femenina (24,2%). En este sentido, un 9% de los bulos no harían referencia a personas, mientras que tan solo una historia falsa versa sobre un hombre y una mujer.

Con respecto a los hombres, cabe mencionar que representan el 85,6% de los representantes políticos; mientras que tan solo uno de los siete artistas analizados es hombre.

Estas cifras pueden deberse, por un lado, a una mayor posición política de liderazgo por parte de los hombres a nivel internacional. Esta situación hace que éstos estén más presentes en la agenda informativa; y en el caso de la mayor presencia de mujeres artistas, sí que se puede afirmar que de las seis analizadas cuatro de ellas habían sido víctimas de propagación de contenido pornográfico falso.

Los tres bulos que no contienen personas tratan sobre una parada de autobús generada a través de inteligencia artificial cuya principal atracción sería que haría a su vez de biblioteca[10]; un robot que está siendo entrenado para participar en la guerra ruso-ucraniana[11]; así como una explosión cerca del Pentágono[12]. El bulo en el que los protagonistas son una pareja de hombre y mujer, trata sobre el abrazo entre una manifestante y un miembro de las fuerzas armadas francesas, durante una de las protestas contra la reforma de las pensiones[13].

Gráfico 5: Número de bulos analizados según el género del sujeto protagonista



Fuente: Elaboración propia

Las temáticas de los bulos (Gráfico 6) se definieron de tal forma que pudieran agrupar el mayor número de *deepfakes* en común, por lo que se evitó la exhaustividad en su definición. No obstante, sí se decidió crear un apartado singular en el que se agrupasen aquellas falsedades relacionadas con políticos y consumo de drogas y alcohol, con el objetivo de diferenciar su número frente al resto.

En este punto, es necesario mencionar que la temática "política" engloba a todas aquellas falsedades relacionadas con representantes políticos, pero que no tengan relación con un consumo de drogas o alcohol. Sin embargo, la categoría "guerra ruso-ucraniana", incluye todos aquellos bulos relacionados con dicho conflicto, contengan políticos en su narración o no. Es decir, se ha decidido privilegiar la guerra ruso-ucraniana frente a la profesión de la persona que aparece en el bulo.

---

[10] *Maldita.es* (2023). No, esta imagen de una parada de autobús en Bélgica no es real: está generada con inteligencia artificial. https://maldita.es/malditobulo/20230417/parada-autobus-belgica-imagen-ia/

[11] *Maldita.es* (2022). No, este vídeo no muestra a un robot real entrenando para el combate: es una animación generada por ordenador.
https://maldita.es/malditobulo/20220624/robot-combate-animacion-generada-ordenador/

[12] *EFE Verifica* (2023). Las imágenes de una explosión cerca del pentágono han sido creadas con inteligencia artificial. https://verifica.efe.com/imagenes-explosion-pentagono-creadas-con-inteligencia-artificial-ia/?utm_medium=Social&utm_source=Twitter#Echobox=1684921757-1

[13] *Maldita.es* (2023). Cómo una imagen creada con inteligencia artificial se viralizó como si fuera una foto real de una manifestación en Francia.
https://maldita.es/malditatecnologia/20230217/imagen-manifestacion-francia-falsa-inteligencia-artificial/

Con respecto a la temática pornográfica, ocurre el mismo fenómeno, puesto que se decide anteponer la materia sobre la que trata el *deepfake* antes que la profesión en la que trabaja el sujeto afectado por el bulo.

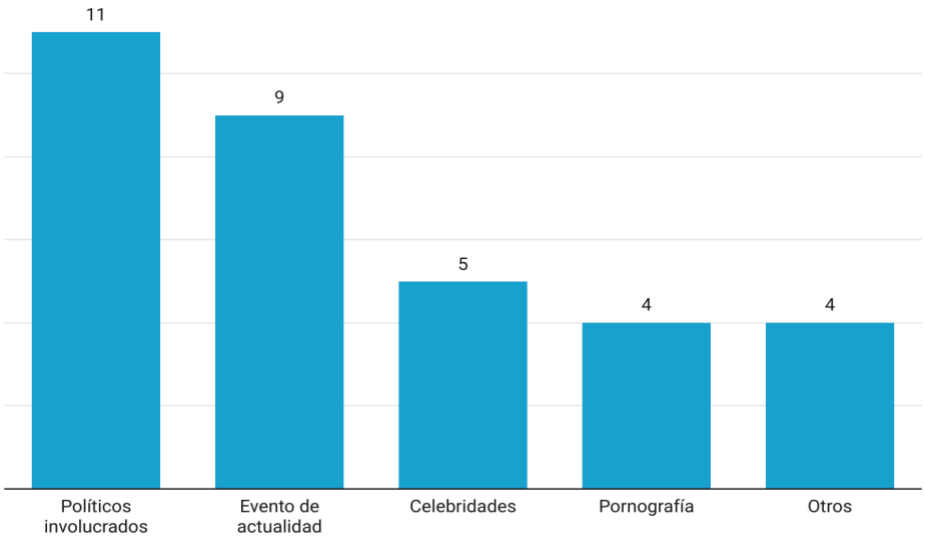
Gráfico 6: Número de bulos analizados según su temática.

Fuente**:** Elaboración propia.

En cualquier caso, en el Gráfico 6 puede observarse que los bulos más frecuentes son aquellos relacionados con la política (48,48%), de los cuáles un 56,25% serían aquellos vídeos difundidos por diferentes usuarios de Internet en los que aparecen representantes políticos en diferentes circunstancias (ruedas de prensa, mítines, discursos, entrevistas…) y cuya voz ha sido ralentizada con el objetivo de hacer creer a los usuarios de Internet que la persona en cuestión había consumido drogas o alcohol previamente a esa intervención pública.

La categoría "otros" contiene aquellos bulos relacionados con tecnología, monarquía, humor, etc., que por su ínfima representación han decidido agruparse de este modo. La guerra ruso-ucraniana y la temática pornográfica serían los protagonistas de un 12,12% de los bulos respectivamente, mientras que los *deepfakes* relacionados con religión siempre involucraban al Papa Francisco.

Gráfico 7. Número de bulos analizados según su motivo de viralidad.



Fuente: Elaboración propia.

Tal y como se ha ido avanzando previamente, la mayoría de los bulos se hacen virales debido a que hacen referencia a personajes públicos. En un amplio número de casos estas personas son representantes políticos que emiten discursos incoherentes dentro del conjunto de su propia narrativa (33,3%), siendo manipulados materiales antiguos para propagar esas difamaciones.

En un número menor de ocasiones (15,5%), ocurre el mismo fenómeno, pero siendo protagonista algún tipo de celebridad o personaje famoso. Además, hay que tener en cuenta que los creadores de bulos deben adaptarse a la agenda mediática con la intención de hacerlos pasar por noticias verdaderas, por lo que partirán desde la base de un evento de actualidad para elaborar esos *deepfakes* (27,3%).

Por último, cabe resaltar un dato realmente significativo, el de los bulos generados con inteligencia artificial cuyo objetivo es hacer creer que ciertas mujeres han participado en vídeos de tipo pornográfico (12,12%). Esta clase de contenido no solo trata de modificar la imagen pública de ciertas mujeres, sino también de invalidarlas a nivel social.

Perspectivas de futuro y conclusiones

Los primeros aprendizajes en torno al uso de la inteligencia artificial en el periodismo y el *fact-checking* nos indican al menos tres perspectivas de estudio: el uso que periodistas y medios le están dando para mejorar su trabajo, cómo altera el modelo de negocio y la manera en la que se comparten los contenidos -incluidos los desinformativos- y cómo puede transformar el seguimiento de historias a través de la automatización de determinados procesos.

Sin embargo, el desarrollo de los *deepfakes* plantea más dudas que certezas en relación a cuestiones como la alfabetización mediática, la legislación y los mecanismos de identificación. Además del uso ilícito relacionado con desnudos y pornografía por parte de las generaciones más jóvenes.

Deepfakes y alfabetización mediática

Desde el MIT, por ejemplo, ya se han puesto en marcha iniciativas como el curso gratuito *Media Literacy in the Age of Deepfakes*[14] con el objetivo de situar los *deepfakes* dentro de una historia más amplia de manipulación de los medios; además de mostrar cómo activistas, artistas, tecnólogos y cineastas están utilizando la inteligencia artificial para una amplia gama de proyectos cívicos.

En este sentido, resulta fundamental empezar a delimitar las distintas iniciativas que se están desarrollando al respecto para que la alfabetización mediática aplicada a la IA establezca categorías, problemáticas y soluciones de forma más rápida (McGowan-Kirsch & Quinlival, 2024), así como identificar y delimitar los límites de su alfabetización (Hwang, Ryu, & Jeong, 2021).

El uso de los *deepfakes* no sólo se presenta como una herramienta de desinformación en procesos electorales, en la actualidad también puede formar parte de las estrategias geopolíticas internacionales o de ataque en batallas culturales determinadas por los ciclos de actualidad. Además, el acceso cada vez más sencillo a esta tecnología y la mejora de calidad de los resultados que ofrece está permitiendo su integración de forma extraordinariamente rápida entre los contenidos que circulan en redes sociales y sistemas de mensajería.

[14] Véase: https://news.mit.edu/2022/fostering-media-literacy-age-deepfakes-0217

Características identificativas de los deepfakes

Al respecto, podemos señalar que las principales características de los *deepfakes* son:

- Utilizan como estrategia el desarrollo de falsos contextos.
- Se sirven principalmente personajes políticos y cantantes y artistas femeninas.
- Se trata de narrativas vinculadas a los ciclos de actualidad.
- Se caracterizan por la ralentización de movimientos.
- Destacan las alteraciones faciales.
- Reproducen la explotación de estereotipos.
- Generalmente son compartidas desde cuentas no oficiales.

Como ocurre con la desinformación tradicional, los falsos contextos -ya sean temporales o geográficos- forman parte de procesos y estrategias que se viralizan de manera rápida. Desde esta perspectiva, y una vez realizado el análisis y propuesto una tipología inicial de análisis en torno a los *deepfakes,* se presentan una serie de recomendaciones para trabajar desde la alfabetización mediática en la identificación de *deepfakes*:

1. Los *deepfakes* se asocian tanto a acontecimientos de actualidad como a personajes que tienen una relevancia pública. Como en el caso de la desinformación tradicional, los *deepfakes* también están determinados por los ciclos de actualidad y por la cercanía o lejanía al contexto informativo en el que se producen.

2. En ellos se pueden identificar características comunes: alteraciones en la imagen propias de las IA (varios dedos, orejas grandes…), modificaciones faciales, ralentización de movimientos, incoherencias en la narrativa propia de un personaje conocido, etc. También hay que tener en cuenta cuestiones como la forma que tienen las manos de sujetar objetos, la desproporción de los zapatos, objetos y lugares que se localizan, etc.

3. Resulta imprescindible una legislación que obligue a las plataformas, aplicaciones y programas como Midjourney o Dall-e que permiten desarrollar y modificar este tipo de contenidos, establecer una marca de agua que los identifique y permita al usuario conocer que esa imagen o vídeo ha sido modificado o creado con inteligencia artificial.

4. Más de la mitad de las verificaciones relacionadas con los *deepfakes* hacen referencia a personalidades políticas. En contextos de intensidad informativa como unas elecciones siempre es necesario dudar de su veracidad, aun cuando tenga objetivos humorísticos.

También resultaría necesario una normativa específica en periodo electoral que comprometiera a los distintos partidos políticos.

5. Los *deepfakes* que tienen como víctimas mujeres muchas veces tienen la intención de silenciarlas en la esfera pública. De hecho, aquellas manipulaciones de tipo pornográfico identificados sólo afectan a mujeres.

6. Muchos de estos posibles *deepfakes* se pueden identificar a través de la búsqueda inversa de imágenes (Google Images, Tin Eye, Bing images, etc.). Es necesario establecer módulos de formación sobre esta cuestión en la educación secundaria pero también sobre las consecuencias jurídicas que puede tener crear y difundir este tipo de contenidos entre menores.

7. Para su identificación, es importante monitorizar las cuentas desde las que se han viralizado las imágenes y valorar la credibilidad del tipo de fuente que está compartiendo ese tipo de contenido, así como la reacción de las cuentas oficiales de las personas implicadas. Al respecto, es importante abrir el debate sobre cómo integrar un módulo de formación sobre esta cuestión en la formación de esa *alfabetización transmedia*.
8. Es creciente el uso de *deepfakes* para desarrollar narrativas y estrategias de comunicación corporativa o para timos y estafas a través de la suplantación de voz o imagen en tiempo real, por lo que se trata de un fenómeno en el que hay que seguir explorando sus consecuencias en diferentes campos y desarrollar campañas de concienciación pública.
9. Los *deepfakes* también tratan de fomentar estereotipos -como los referentes a personalidades políticas- o atacar la imagen de las personas por cuestiones como el género o la raza. Es importante analizar las posibles incoherencias en la narrativa propia de los personajes.
10. El carácter internacional de los *deepfakes* hace necesaria una colaboración más amplia entre organizaciones de fact-checking para compartir información, puesto que muchos de ellos no son identificados en las primeras horas y pueden acabar viralizándose a pesar de desmentidos locales[15].
Sin embargo, queda por analizar con mayor profundidad el papel que pueden tener las redes sociales y las plataformas a la hora de identificar esos *deepfakes* que circulan por sus espacios de difusión.
En este sentido, el uso de la tecnología se presenta siempre por el doble uso de la misma y que genera disfuncionalidades, pero también posibilidades. En el caso de la inteligencia artificial, esta dualidad se ve representada por los *deepfakes* pero también por una automatización y aceleración de los procesos de fact-checking. Como hemos visto, la automatización permitirá un mejor seguimiento de determinados temas, así como aumentar la capacidad para hacer de los datos historias.
Como conclusión nos gustaría señalar que a pesar de las limitaciones que tiene una problemática de reciente aparición como los *deepfakes,* es necesario establecer categorías y mecanismos de identificación tempranos ante la posibilidad de que este fenómeno se normalice.
En cualquier caso, es necesario seguir avanzando en este estudio puesto que su proliferación hace necesaria una investigación continuada de la problemática. Empezando por sus patrones de creación, circulación y respuesta social.
El hecho de que la calidad de los mismos y su realismo sea cada vez sea mayor hace que se necesite de una mayor coordinación entre organizaciones de fact-checking, fuentes oficiales, medios de comunicación y plataformas para impedir su difusión sin etiquetado.
En cualquier caso, en contextos internacionales, tanto los corresponsales como los medios de comunicación se convierten en fundamentales para limitar su viralidad e impedir que se conviertan en armas geopolíticas.

La responsabilidad en materia de alfabetización mediática en el seno de la UE hasta el momento recae, exclusivamente, en los Estados miembros. Por tanto, resultaría pertinente enfocar este tema de forma holística y coordinada, impulsando planes y herramientas a escala europea cuyo uso no se limite a las organizaciones de fact-checking.

[15] Al respecto, y como hemos señalado, entre las ventajas identificadas del uso de la IA en el periodismo y el fact-checking está el reconocimiento de patrones entre grandes cantidades de información que los humanos simplemente no pueden analizar tan rápido o construir gráficos de forma automática que permiten acompañar a las informaciones.

Reconocimientos y financiación

Este trabajo se integra dentro del proyecto IBERIFIER, financiado por la Comisión Europea bajo el acuerdo CEF-TC-2020-2 (European Digital Media Observatory) con la referencia 2020-EU-IA- 0252.

Referencias bibliográficas

Adair, B. (2021, June 28). The lessons of Squash, our groundbreaking automated fact-checking platform. Duke Reporters' Lab. https://reporterslab.org/the-lessons-of-squash-our-groundbreaking-automated-fact-checking-platform/

Adair, B. (2020, February 23). Squash report card: Improvements during State of the Union ... and how humans will make our AI smarter. Duke Reporters' Lab. https://reporterslab.org/squash-report-card-improvements-during-state-of-the-union-and-how-humans-will-make-our-ai-smarter/

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Society, 35(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Babakar, M., & Moy, W. (2016). The State of Automated Factchecking. Full Fact. https://fullfact.org/blog/2016/aug/automated-factchecking/

Boté-Vericad, J.-J., & Vállez, M. (2022). Image and video manipulation: The generation of deepfakes. In P. Freixa, L. Codina, M. Pérez-Montoro, & J. Guallar (Eds.), Visualisations and narratives in digital media. Methods and current trends (pp. 116-127). Barcelona: DigiDoc-EPI. https://doi.org/10.3145/indocs.2022.8

Cerdán Martínez, V., & Padilla Castillo, G. (2019). Historia del «fake» audiovisual: «deepfake» y la mujer en un imaginario falsificado y perverso. Historia y comunicación social, 24(2), 505-520. https://doi.org/10.5209/HICS.66293

De-Lima-Santos, & Salaverría, R. (2021). From Data Journalism to Artificial Intelligence: Challenges Faced by La Nación in Implementing Computer Vision in News Reporting. Palabra-Clave, 24(3), 1–40. DOI: https://doi.org/10.5294/pacla.2021.24.3.7

De-Lima-Santos, M. F., & Ceron, W. (2021). Artificial intelligence in news media: current perceptions and future outlook. Journalism and media, 3(1), 13-26.

Del Castillo, C., & Castro, I. (2023, November 28). Francia, Alemania e Italia piden recortar la regulación de la inteligencia artificial, la gran apuesta de España en la UE. Eldiario.es. https://www.eldiario.es/tecnologia/francia-alemania-e-italia-piden-recortar-regulacion-inteligencia-artificial-gran-apuesta-espana-ue_1_10721155.html

Departamento de Seguridad Nacional. (2023). "Foro contra las Campañas de Desinformación en el ámbito de la Seguridad Nacional - Trabajos 2023". Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática.

Departamento de Seguridad Nacional. (2022). Lucha contra las campañas de desinformación en el ámbito de la seguridad nacional: Propuestas de la sociedad civil. Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática.

European Commission. (2018). Artificial intelligence for Europe. COM (2018) 237 final. DOUE

{SWD(2018)137FINAL}. https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:52018SC0137

European Commission for the efficiency of justice. (2018). European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment. Strasbourg: Council of Europe.

Fernández, P. M. (2018, August 24). Hicimos el primer chequeo en vivo del mundo con transcripción automática. Chequeado. https://chequeado.com/hicimos-el-primer-chequeo-en-vivo-del-mundo-con-transcripcion-automatica/

Fernández, P. M. (2021, April 2). Inteligencia artificial para chequear más rápido y mejor. Chequeado. https://chequeado.com/inteligencia-artificial-para-chequear-mas-rapido-y-mejor/

Funke, D. (2017, November 17). In a step toward automation, Full Fact has built a live fact-checking prototype. Poynter. https://www.poynter.org/fact-checking/2017/in-a-step-toward-automation-full-fact-has-built-a-live-fact-checking-prototype/

Gonzalo, M. (2023, February 8). El cierre de la API de Twitter amenaza la lucha contra la desinformación. Newtral. https://www.newtral.es/cierre-api-twitter-desinformacion/

Graves, L. (2018). Understanding the Promise and Limits of Automated Fact-Checking. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/our-research/understanding-promise-and-limits-automated-fact-checking

Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017). Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada: ACM, 1803–12. https://dl.acm.org/citation.cfm?id=3098131

Hrckova, A., Moro, R., Srba, I., Simko, J., & Bielikova, M. (2022, November 22). Automated, not Automatic: Needs and Practices in European Fact-checking Organizations as a basis for Designing Human-centered AI Systems. arXiv - CS - Human-Computer Interaction DOI:arxiv-2211.12143

Hwang, Y., Ryu, J. Y., & Jeong, S.-H. (Marzo de 2021). Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education. *Cyberpsychology, Behavior, and Social Networking, 24* (3), 188-193. https://doi.org/10.1089/cyber.2020.0174

INCIBE. (2022). Deepfakes. Las apariencias engañan. https://www.incibe.es/aprendeciberseguridad/deepfakes

Lim, G., Perrault, S.T. (2023). Fact Checking Chatbot: A Misinformation Intervention for Instant Messaging Apps and an Analysis of Trust in the Fact Checkers. In: Soon, C. (eds) Mobile Communication and Online Falsehoods in Asia. Mobile Communication in Asia: Local Insights, Global Implications. Springer, Dordrecht. https://doi.org/10.1007/978-94-024-2225-2_11

Maldita.es. (2021). Desinformación en Whatsapp: el chatbot de maldita.es y el atributo reenviado frecuentemente. Maldita.es. https://maldita.es/uploads/public/docs/desinformacion_en_whatsapp_ff.pdf

McGowan-Kirsch, A. M., & Quinlivan, G. V. (2024). Educating emerging citizens: Media literacy as a tool for combating the spread of image-based misinformation. *Communication Teacher, 38*(1), 41-52. https://doi.org/10.1080/17404622.2023.2271548

Moran, R. E., & Shaikh, S. J. (2022). Robots in the news and newsrooms: Unpacking meta-journalistic discourse on the use of artificial intelligence in journalism. Digital journalism, 10(10), 1756-1774.

Morrish, L. (2023). Fact-Checkers Are Scrambling to Fight Disinformation With AI. Wired. https://www.wired.com/story/fact-checkers-ai-chatgpt-misinformation/

Mortera-Franco, P. (2023, November 9). García del Blanco: "Los europeos pueden estar tranquilos, el control de la IA será muy estricto". infoLibre. https://www.infolibre.es/politica/garcia-blanco-europeos-tranquilos-utilizacion-ia-tendra-controles-estrictos_1_1661716.html

Munoriyarwa, A., Chiumbu, S., & Motsaathebe, G. (2023). Artificial intelligence practices in everyday news production: The case of South Africa's mainstream newsrooms. Journalism Practice, 17(7), 1374-1392.

Parlamento Europeo. (2017, February 17). Resolución de 17 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho Civil sobre robótica. P8_TA(2017)0051. 2017, numeral 59.f. https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.pdf

Parlamento Europeo. (2020, October 20). Resolución de 20 de octubre de 2020, con recomendaciones destinadas a la Comisión sobre un régimen de responsabilidad civil en materia de inteligencia artificial. P9_TA(2020)0276. https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_ES.html#title1

Pellicer, M. (2022). Cómo la Inteligencia Artificial está cambiando los medios. MiquelPellicer.com. https://miquelpellicer.com/2022/11/como-la-inteligencia-artificial-esta-cambiando-los-medios-de-comunicacion/

Rocha Jr., D. B., Lins, A. J. da C. C., De Souza, A. C. F., Libório, L. F. de O., Leitão, A. H. de B., & Santos, F. H. S. (2019). VERIFIC.AI application: Automated fact-checking in Brazilian 2018 general elections. Brazilian Journalism Research, 15(3), 514-539. https://doi.org/10.25200/BJR.v15n3.2019.1178

Sánchez-Illán, J.C. (2021). Periodismo frente a desinformación: 2020, el año de la pandemia y de las "fake news". En C. Luena, J.C. Sánchez-Illán, & C. Elías (Eds.), La desinformación en la UE en los tiempos del Covid-19 (pp. 143-149). Tirant lo Blanch.

Sádaba, C., & Salaverría, R. (2023). Combatir la desinformación con alfabetización mediática: análisis de las tendencias en la UE. Revista Latina de Comunicación Social, 81, 17-33. https://www.doi.org/10.4185/RLCS-2023-1552

Shin, J., & Chan-Olmsted, S. (2022). User Perceptions and Trust of Explainable Machine Learning Fake News Detectors. International Journal of Communication, 17(0), Art. 0. https://ijoc.org/index.php/ijoc/article/view/19534

Terrasa, R. (2023, December 9). Así es la primera ley en el mundo sobre inteligencia artificial: "Lo que vendrá en el futuro produce un vértigo espectacular". El Mundo. https://www.elmundo.es/papel/futuro/2023/12/09/65749077e9cf4a36778b45a7.html

Tribunal de Cuentas Europeo. (2021). Informe especial sobre el Impacto de la desinformación en la UE: una cuestión abordada, pero no atajada. https://op.europa.eu/webpub/eca/special-reports/disinformation-9-2021/es/

Ulken, E. (2022). Generative AI brings wrongness at scale. Nieman Lab. https://www.niemanlab.org/2022/12/generative-ai-brings-wrongness-at-scale/

UNESCO. (2021). Recomendación sobre la ética de la Inteligencia Artificial de UNESCO. UNESCO. https://es.unesco.org/fieldoffice/montevideo/EticaInteligenciaArtificial