Deepfakes: el próximo reto en la detección de noticias falsas*

Francisco José García-Ull
Universidad Europea de Valencia
Universidad Miguel Hernández. Instituto Mediterráneo de Estudios de Protocolo
franciscojose.garcia@universidadeuropea.es
fran.garcia@protocoloimep.com

Fecha de presentación: diciembre de 2020
Fecha de aceptación: mayo de 2021
Fecha de publicación: junio de 2021

Cita recomendada: GARCÍA-ULL, F. J. (2021). «*Deepfakes*: el próximo reto en la detección de noticias falsas». *Anàlisi: Quaderns de Comunicació i Cultura*, 64, 103-120. DOI: <https://doi.org/10.5565/rev/analisi.3378>

Resumen

Un *deepfake* o ultrafalso es un vídeo hiperrealista manipulado digitalmente para representar a personas que dicen o hacen cosas que en realidad nunca sucedieron. Estas representaciones sintéticas, generadas mediante técnicas computarizadas basadas en inteligencia artificial (IA), plantean serias amenazas para la privacidad, en un nuevo escenario en el que se incrementan los riesgos derivados de las suplantaciones de identidad. Con la sofisticación de las técnicas para el desarrollo de *deepfakes*, resulta cada vez más complicado detectar si las apariciones públicas o declaraciones de personajes influyentes responden a parámetros de realidad o, por el contrario, son resultado de representaciones ficticias. Este estudio tiene como objetivo plantear un estado de la cuestión a través del análisis de la actualidad académica y mediante una exhaustiva revisión bibliográfica. En el presente trabajo se busca dar respuesta a las cuestiones que planteamos a continuación, que entendemos de interés general, tanto en una vertiente económica y social como en diversas áreas de investigación: ¿qué son los *deepfakes*?, ¿quién los produce y qué tecnología los respalda?, ¿qué oportunidades plantean?, ¿qué riesgos se asocian a estos documentos multimedia?, ¿qué métodos existen para combatir estas falsificaciones? Y enmarcando el estudio en el ámbito de la teoría de la información: ¿se trata de una revolución o de una evolución de las *fake news*? Como sabemos, las noticias falsas influyen en la opinión pública y son efectivas a la hora de apelar a emociones y modificar comportamientos. Podemos asumir que estos nuevos textos audiovisuales serán tremendamente eficaces a la hora de

* Este artículo forma parte del proyecto financiado por el Ministerio de Economía y Competitividad español «Estrategias, agendas y discursos en las cibercampañas electorales: medios de comunicación y ciudadanos» (2017-2020). Referencia CSO2016-77331-C2-1-R.

minar, más si cabe, la credibilidad de los medios digitales, así como de acelerar el ya evidente agotamiento del pensamiento crítico.

Palabras clave: *deepfakes*; noticias falsas; aprendizaje profundo; inteligencia artificial; desinformación

Resum. Deepfakes*: el pròxim repte en la detecció de notícies falses*

Un *deepfake* o hipertrucatge és un vídeo hiperrealista manipulat digitalment per representar persones que diuen o fan coses que en realitat mai no van succeir. Aquestes representacions sintètiques, generades mitjançant tècniques informatitzades basades en intel·ligència artificial (IA), plantegen serioses amenaces per a la privacitat, en un nou escenari en el qual s'incrementen els riscos derivats de les suplantacions d'identitat. Amb la sofisticació de les tècniques per al desenvolupament de *deepfakes*, resulta cada vegada més complicat detectar si les aparicions públiques o declaracions de personatges influents responen a paràmetres reals o, per contra, són resultat de representacions fictícies. Aquest estudi té com a objectiu plantejar un estat de la qüestió a través de l'anàlisi de l'actualitat acadèmica i mitjançant una exhaustiva revisió bibliogràfica. En el present treball es busca donar resposta a les qüestions que plantegem a continuació, que entenem d'interès general, tant en un vessant econòmic i social com en diverses àrees de recerca: què són els *deepfakes*?, qui els produeix i quina tecnologia els dona suport?, quines oportunitats plantegen?, quins riscos s'associen a aquests documents multimèdia?, quins mètodes existeixen per combatre aquestes falsificacions? I emmarcant l'estudi en l'àmbit de la teoria de la informació: es tracta d'una revolució o d'una evolució de les *fake news*? Com sabem, les notícies falses influeixen en l'opinió pública i són efectives a l'hora d'apel·lar emocions i modificar comportaments. Podem assumir que aquests nous textos audiovisuals seran tremendament eficaços a l'hora de minar, més si fos possible, la credibilitat dels mitjans digitals, així com d'accelerar el ja evident esgotament del pensament crític.

Paraules clau: *deepfakes*; notícies falses; aprenentatge profund; intel·ligència artificial; desinformació

Abstract. *DeepFakes: The Next Challenge in Fake News Detection*

A deepfake is a hyper-realistic video, digitally manipulated to represent people saying or doing things that never really happened. With the sophistication of techniques for developing these counterfeits, it is becoming increasingly difficult to detect whether public appearances or statements by influential people respond to parameters of reality or, on the contrary, are the result of fictitious representations. These synthetic documents, generated by computerized techniques based on Artificial Intelligence (AI), pose serious threats to privacy, in a new scenario in which the risks derived from identity theft are increasing. This study aims to advance the state of the art through the analysis of academic news and through an exhaustive literature review, seeking answers to the following questions, which we understand to be of general interest, from both an economic and a social perspective and in various areas of research. What are deepfakes? Who produces them and what technology supports them? What opportunities do they present? What risks are associated with them? What methods exist to combat them? And framing the study in terms of information theory: is this a revolution or an evolution of fake news? As we know, fake news influences public opinion and is effective in appealing to emotions and modifying behaviours. We can assume that these new audiovisual texts will be tremendously effective in undermining, even more if possible, the credibility of digital media, as well as accelerating the already evident exhaustion of critical thinking.

Keywords: deepfakes; fake news; deep learning; artificial intelligence; disinformation

1. Introducción

Las nuevas herramientas basadas en inteligencia artificial (IA) permiten la recreación de representaciones audiovisuales realistas originales que simulan la apariencia y el habla de los seres humanos. Estas representaciones sintéticas se conocen como *deepfake* (palabra compuesta en la que se combinan *deep learning*, o aprendizaje profundo, y *fake*, es decir, falso) y plantean serias amenazas para la privacidad, en un nuevo escenario en el que se incrementan los riesgos derivados de las suplantaciones de identidad. La tendencia demuestra una constatable proliferación de estas técnicas de representación virtual que, unidas a su fácil acceso y usabilidad, hacen posible generar de manera sencilla contenidos multimedia con una falsa apariencia de realidad. Los *deepfakes* pueden generarse a partir de imágenes, audios o vídeos.

Este escenario plantea serios retos que se incorporan a la ya compleja detección de noticias falsas. Los estudios recientes en el ámbito apuntan a que, en efecto, resultará cada vez más complicado detectar técnicamente si las apariciones públicas o declaraciones de personajes influyentes responden a parámetros de realidad o, por el contrario, son resultado de representaciones ficticias.

Los *deepfakes* son cada vez más realistas. Aunque hasta el momento podemos observar algunas características comunes en estas imágenes sintéticas (rostros sin gafas ni barba y que miran directamente a la cámara), el resultado es cada vez más sofisticado. De hecho, si bien los estudios para la identificación de estos vídeos manipulados han resultado eficaces atendiendo a indicadores biométricos (como el número de veces que el protagonista pestañea, diversas expresiones faciales o la composición de la dentadura), la mayoría de especialistas coinciden en que, a medida que se perfeccione la técnica, las diferencias con respecto al sujeto real (o *pristine*) serán cada vez más imperceptibles.

En este artículo explicamos cómo el uso de estas herramientas aplicadas al intercambio, reanimación y manipulación de rostros ha hecho ya saltar algunas alarmas. *Deepfakes* de personalidades como Barack Obama, Donald Trump o Mark Zuckerberg han conseguido acaparar repercusión mediática internacional, mientras que populares actrices o periodistas han sido las primeras víctimas de estas representaciones manipuladas en las que su rostro es insertado en contenidos de carácter pornográfico.

2. Marco teórico

En los últimos años, las noticias falsas se han convertido en un tema que amenaza el discurso público, la sociedad humana y la democracia (Borges, Martins y Calado, 2019; Mackenzie y Bhatt, 2018; Qayyum et al., 2019). En un escenario donde el caos reina en gran parte del ecosistema de información del que dependen las sociedades (Lin, 2019), la información falsa se propaga rápidamente a través de las redes sociales, donde puede impactar a millones de usuarios (Figueira y Oliveira, 2017). Las noticias falsas tienen un importante impacto en la construcción de la realidad por parte del receptor, hasta el punto

de que influyen en su percepción y toma de decisiones, aun a sabiendas de su origen deliberadamente ficticio (Keersmaecker y Roets, 2017). Actualmente, uno de cada cinco usuarios de internet recibe sus noticias a través de YouTube, solo superado por Facebook (Anderson, 2018). Este aumento en la popularidad del vídeo destaca la necesidad de herramientas para confirmar la autenticidad del contenido de los medios y las noticias, ya que las nuevas tecnologías permiten manipulaciones convincentes de vídeos o audios (Anderson, 2018). Dada la facilidad para obtener y difundir información errónea a través de las plataformas de redes sociales, tanto en forma de publicación como en los comentarios (Atasanova et al., 2019), cada vez es más difícil saber en qué confiar, lo que genera consecuencias perjudiciales para la toma de decisiones informadas (Borges et al., 2019; Britt et al., 2019). De hecho, hoy vivimos en lo que algunos autores identifican como un escenario de posverdad, que se caracteriza por la desinformación digital, el sesgo mediático (Hamborg et al., 2018), la generación de información falsa y la distorsión deliberada de la realidad, para manipular creencias y emociones e influir en la opinión pública y en actitudes sociales (Anderson, 2018; Qayyum et al., 2019).

Para identificar los *deepfakes*, debemos comprender las razones de su existencia y la tecnología que los respalda. Sin embargo, dado que los *deepfakes* aparecieron en internet en 2017, la literatura académica sobre el tema todavía es escasa. En este sentido, el presente estudio tiene como objetivo discutir qué son los *deepfakes* y quién los produce, cuáles son los beneficios y las amenazas que introducen, así como mostrar algunos ejemplos de *deepfakes* actuales, para señalar cómo identificarlos y, en su caso, combatirlos. Esta investigación analiza la actualidad académica sobre *deepfakes* y contribuye a la literatura entorno a las noticias falsas y a la desinformación.

Los recientes avances tecnológicos han facilitado la creación de vídeos hiperrealistas que utilizan intercambios de rostros y que dejan pocos rastros de manipulación (Chawla, 2019). Los *deepfakes* son el producto de aplicaciones de IA que fusionan, combinan, reemplazan y superponen imágenes y videoclips para crear vídeos falsos que parecen auténticos (Maras y Alexandrou, 2019). Así, es posible generar, por ejemplo, un vídeo humorístico, pornográfico o político de una persona que dice o hace algo sin su consentimiento (Day, 2019; Fletcher, 2018). El factor que cambia el juego de los *deepfakes* es el alcance, la escala y la sofisticación de la tecnología involucrada, ya que casi cualquier persona con una computadora puede fabricar vídeos falsos que son prácticamente indistinguibles de los medios auténticos (Fletcher, 2018). Si bien los primeros ejemplos de *deepfakes* se centraron en líderes políticos, actrices, comediantes y artistas con rostros entretejidos en vídeos porno (Hasan y Salah, 2019), en el futuro los *deepfakes* probablemente se usarán cada vez más para la pornografía de venganza —*revenge porn*—, el acoso, la evidencia de vídeos falsos en los tribunales, el sabotaje político, la propaganda terrorista, el chantaje, la manipulación del mercado y las noticias falsas (Maras y Alexandrou, 2019).

3. Metodología

Este estudio realiza una revisión bibliográfica, basada en la literatura académica emergente sobre *deepfakes*. En marzo de 2021 se recopilaron un total de 35 artículos académicos centrados en *deepfakes*. Todos estos estudios científicos han sido escritos en inglés y publicados entre 2018 y 2021. Fueron encontrados a través de la búsqueda de Google Scholar, utilizando las palabras clave *deepfake*, *deep fake* y las correspondientes formas plurales. Además, se realizaron búsquedas análogas en los buscadores de las principales editoriales académicas (Springer Nature, Wiley, ElSevier, ACS y Taylor & Francis). El corpus del estudio está compuesto por textos centrados en *deepfakes* que desarrollan sus investigaciones desde distintos prismas o áreas, entre los que se encuentran fundamentalmente las ciencias sociales (periodismo y teoría de la información) y los lenguajes y sistemas informáticos.

4. Resultados

4.1. ¿Qué son los deepfakes*?*

Una combinación de *aprendizaje profundo* y *falso*, los *deepfakes* son vídeos hiperrealistas manipulados digitalmente para representar a personas que dicen y hacen cosas que en realidad nunca dijeron ni sucedieron. Los *deepfakes* se basan en redes neuronales que analizan grandes conjuntos de datos para aprender a imitar las expresiones faciales, los gestos y la voz de una persona. El proceso implica introducir imágenes de dos personas en un algoritmo de aprendizaje profundo para intercambiar caras (Rössler et al., 2018).

Los *deepfakes* salieron a la luz en 2017 cuando un usuario de Reddit publicó vídeos que mostraban a celebridades en situaciones sexuales comprometedoras. Se trata de montajes difíciles de detectar, ya que se basan en grabaciones reales que incluso pueden tener audio con sonido auténtico.

Los *deepfakes* están optimizados para ser compartidos fácilmente en redes sociales, donde las conspiraciones, los rumores y la información errónea se difunden fácilmente. Al mismo tiempo, el escenario de posverdad empuja a las personas a pensar que no pueden confiar en ninguna información a menos que provenga de sus redes sociales, incluidos familiares, amigos cercanos o parientes, y respalda las opiniones que ya tienen. De hecho, muchas personas están abiertas a cualquier cosa que confirme sus puntos de vista existentes, incluso si sospechan que puede ser falsa (Jang y Kim, 2018). Existen aplicaciones gratuitas para generar *deepfakes* de forma sencilla y fácilmente accesibles, que permiten a usuarios con pocas habilidades técnicas y sin ninguna experiencia artística editar vídeos, intercambiar caras, alterar expresiones y sintetizar el habla casi a la perfección.

En cuanto a la tecnología, los *deepfakes* son el producto de las redes generativas antagónicas (RGA), también conocidas como GAN en inglés. Son una clase de algoritmos de IA que se utilizan en el aprendizaje no supervisado, implementados por un sistema de dos redes neuronales que compiten mutua-

mente en una especie de juego de suma cero. Esta técnica, presentada por investigadores de la Universidad de Montreal (Goodfellow et al., 2014), puede generar fotografías que parecen auténticas a observadores humanos. Las RGA se basan en dos redes neuronales artificiales que trabajan juntas para crear medios de aspecto real. Estas dos redes llamadas *el generador* y *el discriminador* se entrenan en el mismo conjunto de datos de imágenes, vídeos o sonidos. Luego, el primero intenta crear nuevas muestras que sean lo suficientemente buenas para engañar a la segunda red, que trabaja para determinar si los nuevos medios que ve son reales. De esa manera, se impulsan entre sí para mejorar.

En el caso de los *deepfake*, el generador de las RGA crea nuevas imágenes basándose en una base de datos de fotografías o vídeos (a mayor número de

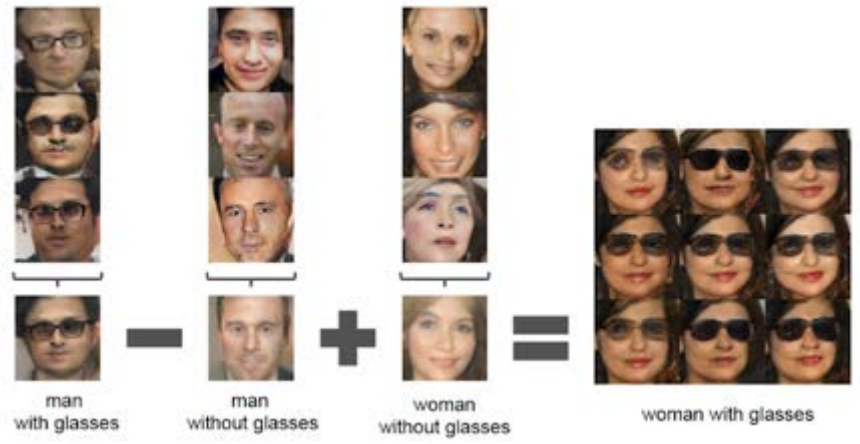
Figura 1. Edmond de Belamy. Primer retrato generado por una IA, subastado en Christie's en Nueva York, 2018

Fuente: Portrait of Edmond de Belamy. Wikipedia.

registros en la base de datos, mayor precisión). Una vez el algoritmo generador crea el nuevo contenido, el algoritmo discriminador realiza un cribado sobre el resultado generado, para delimitar si la imagen o vídeo reúne las características para ser un *deepfake* realista.

Una RGA puede observar miles de fotos de una persona y producir un nuevo retrato que se aproxime a esas fotos sin ser una copia exacta de ninguna de ellas. El resultado es una imagen, vídeo o audio original. En un futuro cercano, las RGA se capacitarán con menos información y podrán intercambiar cabezas, cuerpos enteros y voces. Aunque los *deepfakes* suelen requerir una gran cantidad de imágenes para crear una falsificación realista, los investigadores de la materia están desarrollando técnicas que permiten generar un vídeo falso a partir de una sola fotografía, por ejemplo, una autofoto.

Figura 2. Aritmética vectorial para rostros generados por RGA.



Fuente: Radford, Metz y Chintala (2015).

4.2. ¿Quién produce deepfakes*?*

Podemos distinguir entre al menos cuatro principales actores involucrados en la producción de *deepfakes*, a saber: comunidades de aficionados, actores políticos, delincuentes y estafadores y, por último, actores legítimos, como productores audiovisuales, creadores artísticos (Floridi, 2018) o agencias de publicidad.

En 2017 un usuario de Reddit presentó una colección de *deepfakes* pornográficos de celebridades. En solo unos meses, la comunidad de seguidores que se generó alrededor de *deepfakes* llegaría a los 90.000 miembros.

En estas comunidades, muchos aficionados se centran en *deepfakes* relacionados con la pornografía, mientras que otros colocan a actores famosos en películas en las que nunca aparecieron para producir efectos cómicos. En general, los seguidores tienden a ver los vídeos creados por IA como una

nueva forma de humor en línea y una contribución al desarrollo de dicha tecnología: más como un rompecabezas intelectual, que como una forma de engañar o amenazar a las personas. Sus *deepfakes* están destinados a ser entretenidos, divertidos o centrados en la sátira política, y pueden ayudar a ganar seguidores en las redes sociales. Algunos aficionados pueden estar buscando beneficios personales más concretos, como dar visibilidad al potencial de esta tecnología para exponer sus creaciones y así promocionar su talento.

Sin embargo, también es cierto que existe la amenaza de que actores políticos, *hacktivistas*, agitadores, inferencias de gobiernos extranjeros, etcétera puedan usar *deepfakes* en campañas de desinformación para manipular la opinión pública y debilitar la confianza en las instituciones de un país determinado. En este sentido, los *deepfakes* podrían interferir en unas elecciones o agitar a determinados segmentos para sembrar disturbios civiles.

Los estafadores también están utilizando cada vez más los *deepfakes* con el fin de llevar a cabo la manipulación del mercado y las acciones de una compañía, así como otros delitos financieros. Los delincuentes ya han utilizado audios falsos generados por IA para hacerse pasar por un ejecutivo en el teléfono que solicita una transferencia de efectivo urgente (otra vuelta de tuerca en el conocido fraude del CEO). Existe la tecnología para falsificar videollamadas a tiempo real utilizando imágenes de rostros ya disponibles públicamente en internet.

4.3. Posibles amenazas de los deepfakes

Con relación al objetivo que persiguen y de acuerdo con la bibliografía revisada, la comunidad académica coincide en que los *deepfakes* pueden significar una amenaza para distintos actores sociales, como el sistema político o el empresarial, así como para la ciudadanía en general. En este sentido, estas falsificaciones sintéticas ejercen presión sobre los periodistas que luchan por filtrar noticias reales de las falsas, amenazan la seguridad nacional al difundir propaganda que pueda alterar los resultados de unas elecciones, obstaculizan la confianza de los ciudadanos en la información transmitida por las autoridades y plantean problemas de ciberseguridad tanto para internautas como para organizaciones.

Es muy probable que la industria del periodismo tenga que enfrentarse a un problema masivo de confianza del consumidor debido a los *deepfakes*. Los *deepfakes* representan una amenaza mayor que las noticias falsas tradicionales porque son más difíciles de detectar. La tecnología permite la producción de vídeos de noticias aparentemente legítimos que ponen en riesgo la reputación de los periodistas y los medios de comunicación. Además, ganar la carrera para acceder a las imágenes de vídeo filmadas por el testigo de un incidente puede proporcionar una ventaja competitiva a un medio de comunicación, mientras que el peligro aumenta si las imágenes ofrecidas son falsas.

Los *deepfakes* pueden obstaculizar la alfabetización digital y la confianza de los ciudadanos en la información proporcionada por la autoridad. No obs-

tante, el aspecto más dañino de los *deepfakes* puede que no sea la desinformación *per se*, sino más bien la falta de confianza en las noticias, incluso en formato audiovisual, derivada de un contacto constante con la desinformación. Este aspecto refuerza el fenómeno denominado «apocalipsis de la información» o «apatía por la realidad» (Westerlund, 2019), que derivaría en un agotamiento del pensamiento crítico. Además, las personas pueden incluso descartar imágenes genuinas como falsas, simplemente porque se han arraigado en la noción de que cualquier cosa que no quieran creer debe ser falsa. En otras palabras, la mayor amenaza no reside en el hecho de que el receptor de la información sea engañado, sino en que la información misma pierda toda la credibilidad.

Los problemas de ciberseguridad constituyen otra amenaza impuesta por los *deepfakes*. El mundo empresarial ya ha expresado interés en protegerse contra los fraudes virales, ya que los *deepfakes* podrían usarse para manipular el mercado y las acciones, por ejemplo, al mostrar a un director ejecutivo formulando insultos racistas o misóginos, anunciando una fusión falsa, difundiendo declaraciones falsas sobre pérdidas financieras o retratando a responsables cometiendo prácticas delictivas. Además, los anuncios de productos o pornografía *deepfake* podrían utilizarse para sabotear una marca o como forma de chantaje y extorsión. Además, la tecnología *deepfake* permite la suplantación digital en tiempo real de un ejecutivo, por ejemplo, para pedir a un empleado que realice una transferencia de efectivo urgente o proporcione información confidencial. La tecnología *deepfake* puede incluso crear una identidad fraudulenta y, en vídeos de transmisión en directo, convertir el rostro de un adulto en el rostro de un niño o una persona más joven. Se trata de una posibilidad que genera especial preocupación, dado su potencial uso por parte de redes de pederastia.

4.4. Ejemplos de deepfakes

La mayoría de *deepfakes* de hoy en plataformas sociales como YouTube o Facebook pueden verse como obras artísticas o divertidas inofensivas que utilizan figuras públicas, pero también hay ejemplos del lado oscuro de los *deepfakes*, a saber, el porno de celebridades y de venganza (*revenge porn*), o los intentos de influencia política.

Muchos *deepfakes* se centran en celebridades, políticos y líderes corporativos porque internet está repleto de fotos y vídeos originales de ellos, lo que hace factible la construcción de grandes repositorios de imágenes, necesarios para entrenar un sistema de *deepfake* de IA. La mayoría de estos *deepfakes* son bromas y memes divertidos, con efecto cómico o satírico. Un *deepfake* puede mostrar, por ejemplo, a Sylvester Stallone actuando en películas que nunca ha protagonizado, como *Indiana Jones* o *Terminator 2*.

Figura 3. *Deepfake* en el que Sylvester Stallone protagoniza *Terminator 2*

Fuente: Ctrl Shift Face. YouTube.

Algunos ejemplos interesantes de *deepfakes* incluyen el Museo Dalí en Florida (Estados Unidos), que ha utilizado la tecnología basada en IA para crear montajes que devuelven a la vida a Salvador Dalí (Kwok y Koh, 2020), o el anuncio publicitario de una marca cervecera en el que se muestra una representación sintética de la cantante Lola Flores.

Figura 4. Izquierda: Una representación sintética de Salvador Dalí recibe a los visitantes del Museo Dalí en Florida. Derecha: Un *deepfake* de Lola Flores en el anuncio de la marca Cruzcampo

Fuente: Izquierda, The Dalí Museum. YouTube. Derecha, CruzcampoTV. YouTube.

Sin embargo, también están apareciendo cada vez más ejemplos de *deepfakes* dañinos. Los *deepfakes* permiten la pornografía de celebridades y de venganza, es decir, la pornografía involuntaria con imágenes de celebridades y no celebridades, que se comparte en las redes sociales sin su consentimiento. Así, rostros populares como el de la actriz Scarlett Johansson han aparecido en películas para adultos *deepfakes*, en las que sus caras se han superpuesto a las

de estrellas porno. El resultado es una cosificación de los cuerpos de las mujeres como algo para ser consumido visualmente, que además elude su consentimiento (Wagner y Blewer, 2019). En la escena política, un *deepfake* de 2018 creado por el cineasta de Hollywood Jordan Peele presentaba al expresidente estadounidense Barack Obama discutiendo los peligros de las noticias falsas y burlándose del presidente Donald Trump. Un *deepfake* podría incluso utilizarse para crear falsos recuerdos, dada la maleabilidad de la memoria humana, e impactar directamente en la toma de decisiones de individuos con cargos de alta responsabilidad (Liv y Greenbaum, 2020).

Figura 5. Polémico vídeo en el que los creadores denuncian mediante un *deepfake* de B. Obama los peligros de esta tecnología

Fuente: BuzzVideoFeed. YouTube.

En 2019, un vídeo alterado de la política estadounidense Nancy Pelosi se volvió viral y tuvo un alcance masivo. El vídeo se ralentizó para hacerla parecer intoxicada. En un vídeo *deepfake* de 2018, Donald Trump ofreció consejos a la gente de Bélgica sobre el cambio climático. El vídeo fue creado por un partido político belga —el Partido Socialista Flamenco (SP.A)— con el objetivo de generar un debate social. Hacia el final del vídeo, Trump dice: «Todos sabemos que el cambio climático es falso, al igual que este vídeo». Sin embargo, la última frase no se tradujo en los subtítulos holandeses. El vídeo provocó indignación por la intromisión del presidente estadounidense de un país extranjero en la política climática de Bélgica. En 2019, el Partido Demócrata de los Estados Unidos creó un *deepfake* de su propio presidente, Tom Pérez, para resaltar la amenaza potencial de los *deepfakes* para las elecciones de 2020.

Si bien estos son ejemplos de influencia política limitada, otros *deepfakes* pueden tener un impacto más duradero. En África Central, en 2018, un vídeo del presidente de Gabón, Ali Bongo, a quien se creía en mala salud o

muerto, fue citado como el detonante de un fallido golpe de estado por parte del ejército gabonés. Y en Malasia, un clip viral *deepfake* de la confesión de un hombre de haber tenido relaciones sexuales con un ministro del gabinete local provocó controversia política. También se han utilizado altos ejecutivos para crear *deepfakes*. En junio de 2019, un *deepfake* de alta calidad elaborado por dos artistas británicos con el CEO de Facebook, Mark Zuckerberg, acumuló millones de visitas. El vídeo muestra falsamente a Zuckerberg elogiando a Spectre, una organización ficticia malvada de la serie de James Bond. Con imágenes de noticias, IA y un actor de voz, el vídeo estaba destinado a mostrar cómo se pueden utilizar estas representaciones sintéticas para manipular la realidad.

4.5. Métodos para combatir las falsificaciones profundas

Los artículos revisados sugieren que hay cuatro formas de combatir las falsificaciones profundas: la legislación y la regulación; la concienciación de las políticas corporativas; la educación y capacitación, y la tecnología *antideepfakes*, que incluye detección de *deepfakes*, autenticación de contenido y prevención de *deepfakes*.

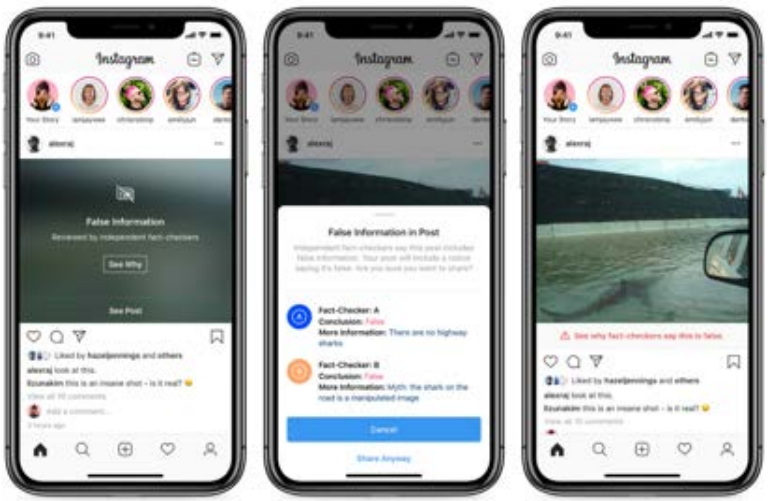
En la actualidad, los *deepfakes* no se tratan específicamente en las leyes civiles o penales, aunque los expertos legales han sugerido adaptar las leyes actuales para cubrir la difamación, la falsificación o la suplantación de identidad. En este sentido, la creciente sofisticación de las tecnologías de IA exige nuevos tipos de leyes y marcos regulatorios. Por ejemplo, los *deepfakes* plantean preocupaciones sobre la privacidad y los derechos de autor, ya que las representaciones visuales de las personas en los vídeos *deepfakes* no son copias exactas de ningún material existente, sino más bien nuevas representaciones generadas por IA. Por lo tanto, los reguladores deben navegar por un panorama legal difícil en torno a las leyes de propiedad y libertad de expresión para regular adecuadamente el uso de esta tecnología.

Por otro lado, el aumento del discurso del odio, las noticias falsas y la desinformación que contaminan las redes sociales (Aldwairi y Alwahedi, 2018) han llevado a algunas empresas a tomar más medidas, como suspender las cuentas de los usuarios e invertir en tecnología de detección más rápida. Reddit y Pornhub han prohibido la pornografía *deepfake* y otra pornografía no consentida, y actúan cuando los usuarios etiquetan y denuncian dicho material.

Facebook pretende prohibir estas falsas representaciones y prevenir su proliferación. Precisamente con este objetivo, la red social ha iniciado una colaboración con la agencia de noticias Reuters destinada a identificar y evitar el efecto viral que podrían causar ciertas imágenes sintéticas. Esta cooperación se enmarca en el programa Facebook Journalism Project, que tiene como objetivo la creación de herramientas de verificación de noticias para combatir la vigente desinformación en la red social. Los algoritmos de Instagram, por otro lado, no recomendarán que las personas vean contenido marcado como falso por los verificadores de hechos (Harrison, 2019; Constine, 2019).

También algunos de los principales medios de comunicación, como *The Wall Street Journal*, *The Washington Post* o Reuters, han tomado medidas para detectar, etiquetar y desmentir información falsa, especialmente *deepfakes* (Vizoso et al., 2021).

Figura 6. *Fact-Checker* de Instagram



Fuente: Constine (2019).

La educación y la formación son cruciales para combatir los *deepfakes*. A pesar de la considerable cobertura mediática y las preocupaciones presentadas por las autoridades, el público aún no ha tenido en cuenta la amenaza de falsificaciones profundas. El estudio de las noticias falsas demuestra que el individuo tiende a pensar que la información falsa afectará con mayor probabilidad a terceras personas que a los miembros de su grupo (Jang y Kim, 2018). Esta sensación ficticia de seguridad crea un ideal caldo de cultivo para la proliferación de paparruchas, bulos y montajes. En este sentido, cobra fundamental importancia la necesidad de concienciar al público sobre el potencial de uso indebido de la IA. Mientras que las representaciones sintéticas brindan a los ciberdelincuentes nuevas herramientas para la ingeniería social, las empresas y organizaciones deben estar en alerta máxima y establecer planes de acción frente a esta latente amenaza. En este sentido, no es de extrañar que numerosos expertos recomienden que en las escuelas se enseñe el pensamiento crítico y la alfabetización digital (Dagdilelis, 2018), ya que estos rasgos contribuyen a la capacidad de los niños para detectar noticias falsas e interactuar de forma más respetuosa entre ellos en línea.

También es importante recordar que la calidad no es un indicador de la autenticidad de un vídeo. Además, las personas deben comprender que, a medida que se desarrolle la tecnología, se requerirán menos fotografías de

rostros reales para crear *deepfakes* y que nadie es inmune. Cualquiera que publique una sola selfi o un vídeo que capture 30 fotogramas por segundo en un sitio de redes sociales corre el riesgo de sufrir una falsificación profunda (Westerlund, 2019). Si bien el mejor método es mantener las fotos y los vídeos fuera de internet, incorporar al vídeo objetos en movimiento o grabar bajo ciertas condiciones de iluminación pueden brindar cierta protección. Las empresas, los gobiernos y las autoridades que utilizan tecnología de reconocimiento facial y almacenan grandes cantidades de datos con fines de seguridad y verificación son actores especialmente sensibles que deben prevenir posibles ataques informáticos y fugas de información.

La tecnología *antideepfake* proporciona quizás el conjunto de herramientas más variado para detectar *deepfakes*, autenticar contenido y evitar que el contenido se utilice para producir representaciones sintéticas.

En la actualidad, se desarrollan en paralelo distintos métodos para detectar *deepfakes* de manera automatizada. Si bien la tecnología basada en IA ya demuestra su eficacia en la detección de noticias falsas (Cybenko y Cybenko, 2018), la gran cantidad de material a analizar representa en el momento un importante reto tecnológico. Por ejemplo, los usuarios cargan 500 horas de contenido por minuto en YouTube. Twitter lucha con ocho millones de cuentas a la semana que intentan difundir contenido manipulado. *Deepfakes* publicados y compartidos en Twitter acumulan millones de visualizaciones (Pérez et al., 2021).

Esto crea enormes desafíos para que las tecnologías revisen todo el material publicado en poco tiempo. Además, los desarrolladores de *deepfakes* tienden a utilizar los resultados de las investigaciones de *deepfakes* publicadas para mejorar su tecnología y sortear nuevos sistemas de detección. Por ejemplo, los investigadores descubrieron que los primeros resultados de *deepfakes* no lograban imitar la velocidad a la que una persona parpadea (Li et al., 2018), mientras que los programas recientes han solucionado la falta de parpadeo o parpadeo antinatural después de que se publicaran los hallazgos.

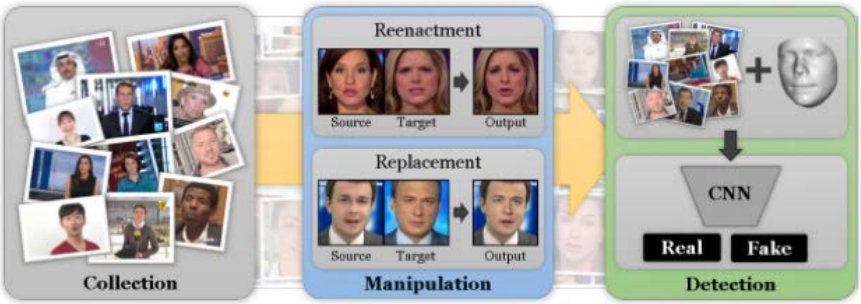
Los expertos en investigación digital forense (Köhn et al., 2006) han sugerido indicadores sutiles para detectar *deepfakes* que incluyen una variedad de imperfecciones en el brillo y la distorsión de la cara; la ondulación en los movimientos de una persona; inconsistencias con el habla y los movimientos de los labios (Korshunov y Marcel, 2019); movimientos anormales de objetos fijos, como un pie de micrófono; inconsistencias en la iluminación, reflejos y sombras; bordes borrosos; ángulos y desenfoque de rasgos faciales; falta de respiración; dirección visual antinatural; frecuencia de parpadeo anormal (Li et al., 2018); falta de rasgos faciales, como un lunar conocido en una mejilla; suavidad y peso de la ropa y el cabello; piel demasiado suave; falta de detalles de cabello y dientes; desalineación en la simetría facial; inconsistencias en los niveles de píxeles, y comportamiento extraño de un individuo que hace algo inverosímil (Westerlund, 2019). Algunos investigadores han demostrado también la eficacia en el análisis computerizado mediante RGA de las principales expresiones faciales (sorpresa, enfado, felicidad, tristeza, miedo, repugnancia y neutralidad) en la detección de *deepfakes* (Anwar et al., 2019). Si

bien cada vez es más difícil para las personas distinguir entre un vídeo real y uno falso, la IA puede ser fundamental para detectar *deepfakes*.

Además, los investigadores que están desarrollando tecnologías RGA pueden diseñar y poner en práctica las salvaguardas adecuadas para que sus tecnologías sean más difíciles de usar indebidamente con fines de desinformación.

Uno de los principales retos tecnológicos actuales persigue una detección automatizada de *deepfakes* cada vez más precisa. En este sentido, las principales plataformas tecnológicas (Google, Facebook o Baidu) están promoviendo competiciones con importantes premios en metálico para animar a la comunidad de desarrolladores a implementar métodos cada vez más fiables en la detección de vídeos falsificados mediante IA. En 2020 se consiguió una precisión en la detección automatizada de *deepfakes* entorno al 80% de acierto. El proyecto «Advanced Deep Learning for Computer Vision», de la Universidad Técnica de Múnich, logró una precisión del 81% a partir de la base de datos Faceforensic++. Por otro lado, el desarrollador Selim Seferbekov resultó ganador del Deepfakes Detection Challenge, con una precisión del 82%.

Figura 7. Funcionamiento de *Faceforensics*++, herramienta para el reconocimiento automatizado de imágenes digitales manipuladas



Fuente: Rössler et al. (2018).

5. Conclusiones

Del mismo modo que afrontamos los retos derivados de la ciberseguridad en general, el primer paso hacia una solución para esta desinformación es comprender el problema y su capacidad para afectarnos. Solo entonces es posible desarrollar e implementar soluciones técnicas que puedan resolver los desafíos. Dicho esto, ninguna de las soluciones tecnológicas puede eliminar por completo el riesgo de *deepfakes*, y el solucionismo tecnológico o la necesidad de buscar soluciones tecnológicas a cada problema, como señala Morozov (2013), pueden incluso desorientar la discusión de preguntas más existenciales sobre por qué existen *deepfakes* y qué otra amenaza puede imponer la IA (Westerlund, 2019). El desafío de la desinformación multimedia habilitada por IA es evolutivo, no revolucionario (Whyte, 2020). No obstante, pode-

mos prever que la democratización de las herramientas basadas en IA producirá en la información efectos estructurales y sistemáticos.

Por lo tanto, y a modo de resumen, asumimos que la estrategia más eficiente para combatir la propagación de los *deepfakes* implica una combinación de avances legales, educativos y tecnológicos y, por supuesto, la implicación de actores gubernamentales, académicos y científicos.

Como sabemos, las noticias falsas, que obedecen a lógicas virales cuando se transmiten mediante redes sociales y servicios de mensajería instantánea, influyen en la opinión pública y son efectivas a la hora de apelar a emociones y modificar comportamientos. En este contexto de posverdad, donde la información se encorseta en píldoras y el titular se prepara como cebo de clics, el incremento de estas técnicas multimedia tiene un efecto catalizador.

No podemos prevenir un escenario donde un *deepfake* decante la balanza en unas elecciones, provoque una crisis bursátil o sea el detonante de una revuelta. Sin embargo, podemos asumir que estas técnicas serán tremendamente eficaces a la hora de minar, más si cabe, la credibilidad de los medios digitales, así como de acelerar el ya evidente agotamiento del pensamiento crítico.

Referencias bibliográficas

ALDWAIRI, M. y ALWAHEDI, A. (2018). «Detecting Fake News in Social Media Networks». *Procedia Computer Science*, 141, 215-222. <https://doi.org/10.1016/j.procs.2018.10.171>

ANDERSON, K. E. (2018). «Getting acquainted with social networks and apps: combating fake news on social media». *Library HiTech News*, 35 (3), 1-6. <https://doi.org/10.1108/LHTN-02-2018-0010>

ANWAR, S.; MILANOVA, M.; ANWER, M. y BANIHIRWE, A. (2019). «Perceptual Judgments to Detect Computer Generated Forged Faces in Social Media». En: SCHWENKER, F. y SCHERER. S. (eds.). *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. MPRSS, 2018. *Lecture Notes in Computer Science*, 11.377. Springer, Cham. <https://doi.org/10.1007/978-3-030-20984-1_4>

ATANASOVA, P.; NAKOV, P.; MÀRQUEZ, L.; BARRÓN-CEDEÑO, A.; KARADZHOV, G.; MIHAYLOVA, T.; MOHTARAMI, M. y GLASS, J. (2019). «Automatic Fact-Checking Using Context and Discourse Information». *Journal of Data and Information Quality*, 11 (3), art. n. 12. <https://doi.org/10.1145/3297722>

BORGES, L.; MARTINS, B. y CALADO, P. (2019). «Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News». *Journal of Data and Information Quality*, 11 (3), art. n.º 14. <https://doi.org/10.1145/3287763>

BRITT, M. A.; ROUET, J.-F.; BLAUM, D. y MILLIS, K. (2019). «A Reasoned Approach to Dealing with Fake News». *Policy Insights from the Behavioral and Brain Sciences*, 6 (1), 94-101. <https://doi.org/10.1177/2372732218814855>

CHAWLA, R. (2019). «Deepfakes: How a pervert shook the world». *International Journal of Advance Research and Development*, 4 (6), 4-8. <http://doi.org/10.22215/timreview/1282>

CONSTINE, J. (2019). «Instagram hides false content behind warnings, except for politicians». *TechCrunch*. Recuperado de <https://techcrunch.com/2019/12/16/instagram-fact-checking>.

CYBENKO, A. K. y CYBENKO, G. (2018). «AI and Fake News». *IEEE Intelligent Systems*, 33 (5), 3-7. <https://doi.org/10.1109/MIS.2018.2877280>

DAGDILELIS, V. (2018). «Preparing teachers for the use of digital technologies in their teaching practice». *Research in Social Sciences and Technology*, 3 (1), 109-121. <http://doi.org/10.46303/ressat.03.01.7>

DAY, C. (2019). «The Future of Misinformation». *Computing in Science & Engineering*, 21 (1), 108-108. <https://doi.org/10.1109/MCSE.2018.2874117>

FIGUEIRA, A. y OLIVEIRA, L. (2017). «The current state of fake news: challenges and opportunities». *Procedia Computer Science*, 121, 817-825. <https://doi.org/10.1016/j.procs.2017.11.106>

FLETCHER, J. (2018). «Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance». *Theatre Journal*, 70 (4), 455-471. ProjectMUSE. <https://doi.org/10.1353/tj.2018.0097>

FLORIDI, L. (2018). «Artificial Intelligence, Deepfakes and a Future of Ectypes». *Philosophy & Technology*, 31 (3), 317-321. <https://doi.org/10.1007/s13347-018-0325-3>

GOODFELLOW, I. J.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A. y BENGIO, Y. (2014). «Generative Adversarial Networks». arXiv:1406.2661.

HAMBORG, F.; DONNAY, K. y GIPP, B. (2018). «Automated identification of media bias in news articles: an interdisciplinary literature review». *International Journal on Digital Libraries*, 20, 391-415. <https://doi.org/10.1007/s00799-018-0261-y>

HARRISON, S. (2019). «Instagram Now Fact-Checks, but Who Will Do the Checking?». *Wired*. Recuperado de <https://www.wired.com/story/instagram-fact-checks-who-will-do-checking/>.

HASAN, H. R. y SALAH, K. (2019). «Combating Deepfake Videos Using Blockchain and Smart Contracts». *IEEE Access*, 7, 41.596-41.606. <https://doi.org/10.1109/ACCESS.2019.2905689>

JANG, S. M. y KIM, J. K. (2018). «Third person effects of fake news: Fake news regulation and media literacy interventions». *Computers in Human Behavior*, 80, 295-302. <https://doi.org/10.1016/j.chb.2017.11.034>

KEERSMAECKER, J. de y ROETS, A. (2017). «Fake news: Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions». *Intelligence*, 65, 107-110. <https://doi.org/10.1016/j.intell.2017.10.005>

KÖHN, M.; OLIVIER, M. S. y ELOFF, J. H. (2006). «Framework for a Digital Forensic Investigation». *ISSA* 1-7.

KORSHUNOV, P. y MARCEL, S. (2019). «Vulnerability assessment and detection of deepfake videos». *International Conference on Biometrics (ICB)*, 1-6. IEEE. <http://doi.org/10.1109/ICB45273.2019.8987375>

KWOK, A. O. y KOH, S. G. (2020). «Deepfake: a social construction of technology perspective». *Current Issues in Tourism*, 1-5. <https://doi.org/10.1080/13683500.2020.1738357>

LI, Y.; CHANG, M. C. y LYU, S. (2018). «In ictu oculi: Exposing AI created fake videos by detecting eye blinking». *IEEE International Workshop on Information Forensics and Security* (WIFS), 1-7. IEEE.

LIN, H. (2019). «The existential threat from cyber-enabled information warfare». *Bulletin of the Atomic Scientists*, 75 (4), 187-196. <https://doi.org/10.1080/00963402.2019.1629574>

LIV, N. y GREENBAUM, D. (2020). «Deep Fakes and Memory Malleability: False Memories in the Service of Fake News». *AJOB Neuroscience*, 11 (2), 96-104. <https://doi.org/10.1080/21507740.2020.1740351>

MACKENZIE, A. y BHATT, I. (2018). «Lies, Bullshit and Fake News: Some Epistemological Concerns». *Postdigital Science and Education*. <https://doi.org/10.1007/s42438-018-0025-4>

MARAS, M. H. y ALEXANDROU, A. (2019). «Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos». *International Journal of Evidence & Proof*, 23 (3), 255-262. <https://doi.org/10.1177/1365712718807226>

MOROZOV, E. (2013). «To save everything, click here: The folly of technological solutionism». *Public Affairs*.

PÉREZ, J.; MESO, K. y MENDIGUREN, T. (2021). «Deepfakes on Twitter: Which Actors Control Their Spread?». *Media and Communication*, 9 (1), 301-312. <http://dx.doi.org/10.17645/mac.v9i1.3433>

QAYYUM, A.; QADIR, J.; JANJUA, M. U. y SHER, F. (2019). «Using Blockchain to Rein in the New Post-Truth World and Check the Spread of Fake News». *IT Professional*, 21 (4), 16-24. <https://doi.org/10.1109/MITP.2019.2910503>

RADFORD, A.; METZ, L. y CHINTALA, S. (2015). «Unsupervised representation learning with deep convolutional generative adversarial networks». *arXiv preprint* arXiv:1511.06434.

RÖSSLER, A.; COZZOLINO, D.; VERDOLIVA, L.; RIESS, C., THIES, J. y NIESSNER, M. (2018). «Faceforensics: A large-scale video dataset for forgery detection in human faces». *arXiv preprint* arXiv:1803.09179.

VIZOSO, A.; VAZ-ÁLVAREZ, M. y LÓPEZ-GARCÍA, X. (2021). «Fighting Deepfakes: Media and Internet Giants' Converging and Diverging Strategies Against Hi-Tech Misinformation». *Media and Communication*, 9 (1), 291-300. <http://dx.doi.org/10.17645/mac.v9i1.3494>

WAGNER, T. L. y BLEWER, A. (2019). «The Word Real Is No Longer Real: Deepfakes, Gender, and the Challenges of AI-Altered Video». *Open Information Science*, 3 (1), 32-46. <https://doi.org/10.1515/opis-2019-0003>

WESTERLUND, M. (2019). «The Emergence of Deepfake Technology: A Review». *Technology Innovation Management Review*, 9 (11), 39-52. <http://doi.org/10.22215/timreview/1282>

WHYTE, C. (2020). «Deepfake news: AI-enabled disinformation as a multi-level public policy challenge». *Journal of Cyber Policy*, 5 (2), 199-217. <https://doi.org/10.1080/23738871.2020.1797135>