EDICIONES COMPLUTENSE

Derecom
e-ISSN: 1988-2629

ARTÍCULOS

La metamorfosis de la verdad: *Deepfakes* y el desafío de la autenticidad en la sociedad digital

Ana Alzaga Gallo
Universidad Rey Juan Carlos (España)

https://dx.doi.org/10.5209/dere.102344 Recibido: 10/01/2025 • Revisado: 17/02/2025 • Aceptado: 05/03/2025

ES **Resumen.** Insertos en la revolución tecnológica actual, la Inteligencia Artificial (IA) y los *deepfakes* han emergido como herramientas poderosas, que no solo facilitan la creación de contenido, sino que también tienen el potencial de distorsionar la realidad de forma jamás vista. Los *deepfakes*, que combinan técnicas avanzadas de aprendizaje profundo, con la capacidad de generar imágenes, vídeos y audios falsos, pueden representar una amenaza significativa para la veracidad de la información, así como para los derechos fundamentales de las personas. Este artículo busca proporcionar una comprensión profunda de cómo los *deepfakes* están redefiniendo el panorama de la información. A través de una exhaustiva revisión bibliográfica, se examinará el concepto de *deepfake* y las tecnologías subyacentes, su impacto social y ético, y las posibles medidas de detección y mitigación. En última instancia, se discutirán sus implicaciones legales y se presentarán estadísticas y ejemplos recientes que ilustran la magnitud de este fenómeno. Entre las principales conclusiones fruto de la revisión realizada en este artículo, se subraya la importancia de abordar los desafíos que presentan los *deepfakes* desde múltiples frentes, incluyendo la tecnología, la legislación y la educación, para proteger la integridad de la información en la era digital.
Palabras clave. Deepfake, Inteligencia Artificial, ciberseguridad, protección de datos, desinformación, privacidad, RIA.

ENG The Metamorphosis of Truth: Deepfakes and the challenge of authenticity in digital society

ENG **Abstract.** Embedded in the current technological revolution, Artificial Intelligence (AI) and deepfakes have emerged as powerful tools, not only facilitating the creation of content but also having the potential to distort reality like never before. Deepfakes, which combine advanced deep learning techniques with the ability to generate fake images, videos, and audio, can pose a significant threat to the veracity of information as well as the fundamental rights of individuals. This article aims to provide a deep understanding of how deepfakes are redefining the information landscape. Through a comprehensive literature review, the concept of deepfake and the underlying technologies, their social and ethical impact, and potential detection and mitigation measures will be examined. Ultimately, their legal implications will be discussed, and recent statistics and examples illustrating the magnitude of this phenomenon will be presented. Among the main conclusions drawn from the review conducted in this article, the importance of addressing the challenges posed by deepfakes from multiple fronts, including technology, legislation, and education, to protect the integrity of information in the digital age is highlighted.
Keywords. Deepfake, Artificial Intelligence, cybersecurity, data protection, disinformation, privacy, AI Regulation (RIA).

Sumario. 1. Introducción. 2. Objetivos y metodología. 3. Concepto y evolución histórica. 4. Creación de Deepfakes. Tecnologías subyacentes: CNN, GANs y VAEs. 5. Estudio de casos de Deepfakes relevantes. 6. Tendencias clave, prevalencia e impacto de los Deepfakes en el panorama audiovisual actual. 7. Deepfakes y su uso malintencionado. Análisis de la amenaza. 8. La detección de Deepfakes: el desafío. 9. Implicaciones y retos legales de los Deepfakes. 10. Conclusiones. 11. Limitaciones. 12. Prospectiva. 13. Referencias.

Cómo citar: Alzaga Gallo, A. (2025). La metamorfosis de la verdad: *Deepfakes* y el desafío de la autenticidad en la sociedad digital. *Derecom* 38(1), 45-57. https://dx.doi.org/10.5209/dere.102344

1. Introducción

En la era digital actual, la información fluye a una velocidad sin precedentes, transformando la manera en que se consume y se comparte contenido. Sin embargo, esta revolución tecnológica también ha generado nuevos desafíos, entre los cuales, destaca la desinformación a través de los *deepfakes,* cuyo concepto, características y proceso de creación serán expuestos con profundidad en este artículo.

El análisis se centrará en cómo estas tecnologías permiten la creación rápida y sencilla de contenido falso con apariencia real, lo cual podría llegar a erosionar la confianza de la opinión pública.

Este artículo analizará esta transformación radical de la información y a continuación, se estudiará el impacto social y ético de los *deepfakes,* centrándose en la amenaza que pueden representar para la privacidad y la seguridad, especialmente en lo relacionado con la autenticación biométrica.

Se abordarán también los desafíos que se plantean en cuanto a su detección y mitigación, subrayando la necesidad de desarrollar algoritmos más sofisticados y métodos de verificación eficaces. Se ha decidido no abundar en exceso en cuestiones técnicas específicas con el fin de mantener el enfoque en los aspectos más accesibles del tema. Esta elección permite que el contenido sea comprensible, sin perder de vista los puntos clave y las implicaciones más importantes. También se analizarán las implicaciones legales y la necesidad de regulación, destacando la urgencia de establecer marcos legales que protejan la privacidad y la seguridad de las personas.

Igualmente, se presentarán estadísticas y ejemplos notables de *deepfakes* reconocidos con el fin de ilustrar la magnitud y la prevalencia de este fenómeno, así como su impacto en el panorama audiovisual actual.

Por último, se subrayará la importancia de la educación y concienciación pública, resaltando la necesidad de campañas educativas específicas para aumentar el conocimiento sobre los *deepfakes* y sus riesgos.

Tras el análisis realizado en este artículo, a modo de cierre, se presentarán las principales conclusiones de este, que subrayan la creciente amenaza que representan los *deepfakes* y la necesidad urgente de desarrollar tecnologías de detección y marcos legales para mitigar sus riesgos.

2. Objetivos y metodología

Los objetivos primordiales de este artículo se centran en tres líneas fundamentales. El objetivo prioritario es evaluar cómo los *deepfakes* están transformando el panorama de la comunicación y cómo afectan a la percepción pública y la confianza en la información, ya que comprender el impacto de los *deepfakes* es crucial para el posterior desarrollo de estrategias efectivas de atenuación y concienciación. En segundo lugar, es esencial un conocimiento profundo de las tecnologías que permiten la creación de *deepfakes*, como las Redes Neuronales Convolucionales (CNN), las Redes Generativas Adversariales (GANs) y los Autoencoders Variacionales (VAEs). La comprensión del funcionamiento de estas tecnologías es esencial para desarrollar herramientas de detección y comprender las capacidades y limitaciones de los *deepfakes*. Un tercer objetivo consiste en evaluar las implicaciones sociales y éticas de los *deepfakes*, incluyendo su impacto en la privacidad y la seguridad, dado que esto es fundamental para el establecimiento de marcos legales y normativos que protejan a los individuos y a la sociedad.

Este conjunto de objetivos proporciona una base sólida para abordar el tema de los *deepfakes* y los desafíos que presentan desde múltiples perspectivas, incluyendo la tecnológica, la ética y la legislativa. En cuanto a las hipótesis de trabajo, la hipótesis central de este artículo es que la Inteligencia Artificial (IA) y los *deepfakes* están transformando radicalmente el fenómeno de la circulación de la información errónea, no solo en términos de la facilidad y velocidad con la que se puede crear contenido falso, sino también en la profundidad del impacto social y ético que estos avances tecnológicos están teniendo en la sociedad contemporánea. Se postula que la capacidad de los *deepfakes* para generar imágenes, vídeos y audios falsos con un alto grado de realismo está erosionando la confianza pública en la veracidad de la información y socavando la integridad de las instituciones democráticas.

Una segunda hipótesis de partida es la de la seria amenaza y el alto riesgo que estas creaciones sintéticas suponen para la privacidad de las personas y los retos para la seguridad en lo relacionado con la autenticación biométrica.

En último lugar, la posibilidad de creación de vídeos y audios hiperrealistas genera numerosas dudas respecto a la posibilidad de detección de la manipulación y la identificación de los deepfakes por parte de los individuos.

En cuanto a la metodología, el presente artículo lleva a cabo una revisión bibliográfica enmarcada dentro de un enfoque metodológico de carácter cualitativo, basado en el análisis exhaustivo de fuentes de literatura académica relevantes, primarias y secundarias, tanto nacionales como internacionales, de autores expertos y acreditados que versan sobre el tema en revisión. 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 en sus ramas de teoría de la información y periodismo, junto con la tecnología, lenguajes y sistemas involucrados.

Para todo ello, se han utilizado recursos tales como bases de datos, bibliotecas especializadas y repositorios de artículos científicos, respetando los más altos estándares de ética y rigor científico. Estos estudios científicos fueron encontrados a través de búsquedas exactas en las principales fuentes académicas *(Eric, Bielefeld Academia Search Engine (BASE), Science Research, Royal Anthropological Institute, American Statistical Association, IEEE, Index Copernicus International)* utilizando las palabras clave "*deepfake",* "*deep fake",* y sus correspondientes formas plurales. La búsqueda se ha acotado temporalmente al periodo comprendido entre los años 2020 y 2024, obteniendo un total de 5.348 resultados. Estos resultados se han delimitado por tipo de documento, prefiriendo documentos Open Access, con lo que los resultados se redujeron

a 638, los cuales fueron filtrados por idioma, inglés (220) y español (9) revisando un total de 229 publicaciones cuyo contenido gira alrededor de la problemática de la desinformación y los *deepfakes* generados con IA. Se ha llevado a cabo un importante trabajo de síntesis de la información debido a la vasta cantidad de datos disponibles sobre el tema. Este proceso ha implicado la recopilación, evaluación y condensación de múltiples fuentes para proporcionar una visión coherente y comprensible.

3. Concepto y evolución histórica

En los últimos tiempos el uso de herramientas basadas en Inteligencia Artificial (IA) ha facilitado la proliferación de representaciones audiovisuales de carácter absolutamente realista, que muestran a seres humanos realizando acciones que nunca sucedieron en realidad. Contenidos en forma de vídeos o audios que pueden representar a personas reales o ficticias. Se trata de representaciones artificiales, hiperrealistas manipuladas digitalmente (García-Ull, 2021) denominados *deepfakes.* El término *deepfake* es una combinación de *deep learning* (aprendizaje profundo) y *fake* (falso), y se refiere a la capacidad de las redes neuronales profundas para generar contenido altamente realista. Estas redes examinan inmensas colecciones de datos con el objetivo de conseguir un aprendizaje profundo que les permita con posterioridad emular con perfecto realismo (Rössler y otros, 2018) tanto las expresiones faciales, como los gestos o la voz de las personas.

No obstante, el término *deepfake* surgió como tal en 2017, dentro de una plataforma en línea denominada *Reddit*, en la que un usuario anónimo comenzó a difundir un vídeo manipulado utilizando tecnología rudimentaria de intercambio de rostros. Este usuario publicó este contenido baja el alias de "*Deepfakes*", con lo que a medida que este tipo de vídeos se popularizó, el término *deepfake* comenzó a utilizarse para denominar de este modo al contenido audiovisual artificial que simula de forma realista la apariencia o voz de personas humanas. Este primer vídeo tenía carácter pornográfico y mostraba escenas en las que aparecía la actriz protagonista de *Wonder Woman*, (Jenkins, 2017) Gal Gadot, tomando como base un vídeo pornográfico ya existente, en el que se mapeó la cara de Gadot sobre la de la actriz del vídeo original. A partir de ese momento, comenzaron su evolución imparable hasta lograr una apariencia cada vez más realista

Los *deepfakes* se presentan en diversas modalidades como las siguientes:

1. Imágenes: estos *deepfakes* muestran imágenes de caras estáticas generadas con técnicas diversas (Sohl-Dickstein y otros, 2015) que incluyen la generación desde cero, *morphing*, con la mezcla de imágenes de características similares o intercambios de unos rostros por otros (Goodfellow et al., 2014).
2. Audios: esta modalidad está relacionada con la creación de audios que suenan como las voces originales de las personas.
3. Vídeo: los *deepfakes* de vídeo (Damer et al., 2018) son producto de la manipulación de imágenes y audio para hacer creer al espectador que la persona está haciendo o diciendo algo que en realidad no hizo (Zhang, 2022).

Las aplicaciones más destacadas de esta tecnología son:

- Intercambio de rostros: se utiliza una gran cantidad de imágenes del rostro de una persona para entrenar el modelo, que luego reemplaza su rostro en vídeos existentes. El producto final es la sustitución del rostro de una persona por otra.
- Creación de escenarios hiperrealistas: cambiando las condiciones de iluminación, recopilando datos en diversas posiciones y circunstancias, se consigue la generación de diferentes escenarios cuya falsedad es difícilmente detectable.
- Resolución de las imágenes: las imágenes en alta resolución permiten al modelo captar detalles finos, como textura de la piel, arrugas y movimientos sutiles.
- Recreación de voz: entrenando modelos con grabaciones de la voz de una persona para generar audio sintético que coincida con su patrón de habla. De este modo se pueden crear declaraciones o discursos falsos que parecen absolutamente auténticos.

La evolución histórica de los *deepfakes* es un aspecto crucial para lograr entender el impacto y las implicaciones de esta tecnología en la actualidad. La manipulación de imágenes (Ajder et al., 2019) no es un concepto nuevo, muy al contrario, se considera que los orígenes de los *deepfakes* se pueden remontar al siglo XIX, vinculados a los inicios de la fotografía (Fineman, 2012) como base de las prácticas digitales actuales. En ese siglo se comenzó a utilizar técnicas de retoque y tratamiento de las imágenes, observándose también antecedentes de las modernas preocupaciones sobre las imágenes digitales, con la existencia de tensiones entre los partidarios de la fotografía como captura objetiva de la realidad y los defensores de la modificación de las imágenes con fines artísticos (Kaplan, 2005).

Ya en estos primeros años de la fotografía surgieron intentos de modificar las imágenes originalmente captadas (Brugioni, 1999), con la utilización de diversas técnicas de manipulación. Entre ellas, cabe mencionar la del retoque manual, con el uso de lápices o pinceles para modificar los negativos y positivos fotográficos. Asimismo, también se aplicaban emulsiones y tintes o técnicas más rudimentarias, como el simple raspado de las imágenes para añadir o eliminar detalles de las escenas, modificar los rasgos faciales o eliminar imperfecciones de los protagonistas (Rosenblum, 2008). Igualmente, también se recurría a la exposición múltiple y la superposición, combinando varias imágenes en una sola fotografía, de modo que se incluían elementos no presentes en la toma original (Batchen, 1999).

En este sentido, uno de los ejemplos más notables de manipulación fotográfica que generó gran interés y polémica, fue el retrato de Abraham Lincoln (1865), en el que Alexander Ritchie, tomó la fotografía de la cabeza de Lincoln realizada por el fotógrafo Antony Berger para el billete de cinco dó-

lares y la superpuso al cuerpo del también político John C. Calhoun, firme defensor de la esclavitud (Ostendorf, 1998). Por su parte, William H. Mumler demostró el poder de la manipulación fotográfica en 1860 a través de la técnica de la doble exposición de imágenes, creando la ilusión de presencias fantasmales (Coates, 2018), asegurando ser capaz de captar imágenes de espíritus junto a personas vivas (Leeder, 2015).

Este temprano ejemplo de manipulación visual se relaciona estrechamente con los antecedentes históricos de los *deepfakes.* Todo ello gracias al pistoletazo de salida iniciado por el trabajo *A Trip to the Moon* (Méliès, 1902), que dio vida a la visión del mundo de H.G. Wells en una película de ciencia ficción, preconizando el uso de efectos especiales en el cine. Este trabajo puede considerarse un antecedente histórico de los *deepfakes*, ya que sentó las bases para la manipulación visual en medios audiovisuales. Al respecto, cabe mencionar igualmente, por su íntima relación con la prehistoria de la tecnología y los efectos visuales, la técnica del *Stop Motion* como una de las formas más antiguas de animación. Esta técnica, que consiste en crear la ilusión de movimiento en objetos o modelos fotografiándolos después de moverlos muy ligeramente, sigue utilizándose en la actualidad. Al reproducir las imágenes tomadas en secuencia rápida, los objetos parecían moverse. Los trabajos de Méliès fueron fundamentales para el desarrollo de esta técnica y de los efectos especiales iniciales. Otros hitos más recientes en la historia de los *deepfakes* y su evolución histórica están marcados, sin duda por la transición al mundo digital con la aparición de las computadoras en la década de los 80. Del mismo modo, el lanzamiento del software especializado que revolucionó la edición digital de imágenes, *Adobe Photoshop* (1989) el programa que hizo accesible la manipulación digital de imágenes a una amplia audiencia, permitiendo ediciones de imágenes complejas y precisas.

En este punto, es imprescindible realizar una mención a otro antecedente de los *deepfakes* en su modalidad de vídeo, haciendo referencia a los efectos especiales en el cine. Un ejemplo significativo de estos primitivos efectos especiales, lo constituyen las imágenes generadas por ordenador (CGI *Computer Generated Images*), que transformaron la industria cinematográfica facilitando la creación de personajes y mundos no existentes en realidad.

La película *Westworld* (Crichton, 1973) es un ejemplo pionero en el uso de efectos especiales a través de un rudimentario sistema de procesamiento digital de imágenes (American Cinematographer, 1973). Esta película no solo destaca por su innovador uso de la tecnología, sino que también enlaza curiosamente con otra obra del mismo autor, Michael Crichton. Años más tarde, Crichton escribió la novela en la que se basó *Jurassic Park* (Spielberg, 1993), donde los dinosaurios y los espacios realistas tomaron vida gracias a la creación íntegra con CGI. Así, *Westworld* y *Jurassic Park* representan hitos importantes en la evolución de los efectos especiales en el cine, mostrando la progresión desde los primeros sistemas digitales hasta el CGI avanzado.

4. Creación de Deepfakes. Tecnologías subyacentes: CNN, GANs y VAEs

Una tecnología fundamental para la creación de deepfakes son las Redes Neuronales Convolucionales (CNN). Estas redes procesan imágenes y vídeos para generar nuevas imágenes realistas. Las CNN (Roy et al. 2022) son una clase de redes neuronales artificiales que se utilizan principalmente para el reconocimiento y procesamiento de imágenes. Están inspiradas en la forma en que el cerebro humano procesa la información visual. La estructura típica de una CNN (Mallet et al. 2023) tiene varias capas, cada una con una función específica:

1. Capa de entrada: recibe la imagen en bruto.
2. Capas convolucionales: aplican filtros a la imagen –llamados *kernels*– para detectar características como bordes, texturas y patrones.
3. Capas de agrupación: Reducen la dimensionalidad de los datos –*pooling*– manteniendo las características más relevantes.
4. Capas completamente conectadas: Después de las capas convolucionales y de agrupación, los datos se aplanan y se pasan a través de una o más capas completamente conectadas que realizan la clasificación final.

Las CNN favorecen la creación de *deepfakes* debido a su ayuda en el reconocimiento de objetos en imágenes, a la segmentación y delineación de regiones específicas dentro de las mismas y clasificación de objetos dentro una imagen, así como la generación de nuevas imágenes realistas a partir de los datos existentes.

La creación de *deepfakes* implica el uso de IA y aprendizaje profundo, utilizando redes neuronales como las GANs (redes generativas adversariales) y autoencoders variacionales (VAEs), además de algoritmos de reconocimiento facial. Los VAEs ayudan con los intercambios de rostros realistas, al comprimir y reconstruir caras, creando *deepfakes* donde el rostro objetivo imita las expresiones y movimientos del origen. Las GANs cuentan con un generador que crea datos y un discriminador que detecta falsificaciones. Este proceso continúa hasta que los datos falsos son difíciles de distinguir de los datos reales. Estos grandes conjuntos de datos entrenan los *deepfakes*. Cuanto más completo y de mayor calidad sea el conjunto de datos, más auténticos deberían aparecer los resultados. Para los vídeos, el nuevo material se renderiza cuadro por cuadro para que los movimientos parezcan naturales.

El posprocesamiento puede involucrar ajustes de autenticidad para hacer coincidir colores, iluminación y otros detalles. Las Redes Generativas Adversariales (GANs) (Goodfellow, 2014) son una arquitectura de aprendizaje con el propósito principal de generar datos artificiales (como imágenes, vídeos o audios) tan realistas que sean prácticamente indistinguibles de los datos reales. Las GANs están compuestas por dos redes neuronales (Radford et al., 2015) que compiten entre sí, en un proceso que mejora continuamente la calidad del contenido generado: generador (*generator)* y discriminador (*discriminator*). La primera, crea datos artificiales a partir de ruido o entradas aleatorias y la segunda tiene

como función evaluar la autenticidad de los datos, distinguiendo entre los datos reales y generados. Estas dos redes se entrenan de manera simultánea y competitiva, de forma que el generador crea datos falsos y el discriminador intenta clasificarlos correctamente. A medida que el generador mejora, crea datos más realistas, obligando al discriminador a afinar su capacidad para detectar falsificaciones.

Las GANs tienen tres aplicaciones fundamentales en la creación de *deepfakes*: el intercambio de rostros, la recreación de expresiones faciales y la generación de clips de audio y vídeo ficticios que parecen auténticos (Karras, 2019). Adicionalmente, como se comentó con anterioridad, en la trastienda de los *deepfakes* se encuentran los *Variational Autoencoders* (VAEs), un tipo de red neuronal que se utiliza para aprender representaciones eficientes de datos de alta dimensión. Funcionan codificando datos de entrada en un espacio de menor dimensión y que se decodifican para reconstruir la entrada original (Dagar & Vishwakarma, 2022). La diferencia clave con otros autoencoders es que los VAEs (Nguyen et al., 2022) introducen una distribución probabilística en el espacio latente, lo que permite generar nuevas muestras de datos que son similares a los datos de entrenamiento. Mediante el proceso de codificación los VAEs convierten datos de alta dimensión, como imágenes, en una representación de menor dimensión, conocida como espacio latente. Esta representación captura las características más importantes de los datos de entrada, mientras que, con la decodificación, el espacio latente se utiliza para generar datos de alta dimensión, produciendo imágenes o vídeos que se parecen a los datos originales. Los VAEs utilizan distribuciones probabilísticas en el espacio latente, lo que permite generar nuevas muestras de datos. Esta regularización ayuda a asegurar que las muestras generadas sean coherentes y realistas.

A pesar de la compleja tecnología subyacente a los *deepfakes*, crear un video *deepfake* simple no es difícil. Se pueden usar teléfonos inteligentes o tabletas para crear *deepfakes* con aplicaciones como *MSQRD*, *Zao* o *FaceApp*. Si las fuentes de imagen, vídeo o audio son buenos, se podría tener un *deepfake* de alta calidad en menos de diez minutos si se emplean las aplicaciones adecuadas.

En concreto, el vídeo que dio origen al término *deepfake* fue realizado en un equipo doméstico, con una herramienta denominada *TensorFlow*. Se trataba de una biblioteca de *software* de código abierto para el aprendizaje automático, desarrollada por el equipo de *Google Brain,* que vio la luz en 2015 y se convirtió en una herramienta popular para la investigación y la implementación de modelos de aprendizaje profundo, que sigue utilizándose en la actualidad. Entre sus características principales *TensorFlow* permite la creación de modelos tanto en servidores como en dispositivos móviles, lo que le dota de una gran flexibilidad. Igualmente admite varios lenguajes de programación, como Python, C++ o Java. Por último, *TensorFlow* ofrece herramientas para visualizar y depurar modelos de aprendizaje profundo. *TensorFlow* fue clave en el desarrollo de deepfakes, gracias a su capacidad para procesar grandes cantidades de datos y entrenar modelos complejos facilitando la creación de deepfakes de alta calidad.

En la actualidad *TensorFlow* se sigue utilizando y juega un papel protagonista en la elaboración de *deepfakes* en lo que respecta principalmente a tres aspectos: en primer lugar, en lo referente al entrenamiento de modelos, permitiendo entrenar modelos de aprendizaje profundo utilizando grandes conjuntos de datos. Estos modelos pueden aprender a reconocer y replicar las características faciales y voces de una persona. En segundo lugar, en la generación de imágenes y vídeos: Una vez que el modelo está entrenado, se puede utilizar para generar imágenes o vídeos falsos. Esto se hace alimentando al modelo con nuevas imágenes o vídeos y ajustando los parámetros para crear una versión falsificada. Por último, con la optimización y ajuste de los modelos, lo que permite mejorar la calidad y realismo de los *deepfakes*.

Otro aspecto reseñable en la creación de *deepfakes* es el del entrenamiento, siendo necesario enfatizar la importancia de los datos durante el mismo, ya que los modelos de *deepfake* se entrenan utilizando grandes volúmenes de datos de video y audio correspondientes a la persona que se desea imitar. La cantidad de datos disponibles es un factor crítico, que determina la calidad y el realismo del resultado final. Cuantos más datos se empleen, más precisa será la simulación, ya que los algoritmos de Inteligencia Artificial (IA) podrán aprender y replicar de manera efectiva las características distintivas de la persona imitada, tales como sus expresiones faciales, movimientos, voz y gestos. Estos datos son fundamentales para que los modelos de IA capturen las sutilezas y matices propios de la persona objetivo, permitiendo la generación de contenido visual y auditivamente indistinguible del material auténtico.

El proceso de entrenamiento se desarrolla en las siguientes fases:

1. Recolección de datos:
   - Imágenes y vídeos: se recopilan imágenes y vídeos de la persona objetivo desde diferentes ángulos, en diversas condiciones de iluminación y con varias expresiones faciales.
   - Audios: se recopilan grabaciones de la voz de la persona para entrenar los modelos de síntesis de voz.
2. Preprocesamiento:
   - Normalización: las imágenes y vídeos se preprocesan para normalizar el tamaño, el formato y la iluminación, asegurando una base de datos coherente.
   - Etiquetado: las características faciales clave, como ojos, nariz, boca y contorno del rostro, se etiquetan para guiar al algoritmo en su aprendizaje.
3. Entrenamiento del modelo:
   - Redes Neuronales Convolucionales (CNN): estas redes analizan y procesan las imágenes para extraer sus características relevantes. Las CNN son extremadamente eficaces en el reconocimiento de patrones en datos visuales.
   - Redes Generativas Antagónicas (GANs): en este proceso, un generador crea imá-

genes falsas y un discriminador evalúa su autenticidad. La competición entre ambas redes mejora continuamente la calidad de las imágenes generadas.

4. Optimización continua:
   - Retroalimentación: los errores detectados por el discriminador se utilizan para ajustar y mejorar el generador.
   - Iteraciones: el proceso se repite durante múltiples iteraciones hasta que las imágenes artificiales sean casi indistinguibles de las reales.

El seguimiento de este proceso de entrenamiento asegura que los modelos de *deepfake* alcancen un alto grado de realismo, permitiendo la generación de contenidos audiovisuales que son difícilmente diferenciables de los originales.

5. Estudio de casos de deepfakes relevantes

El presidente Richard Nixon, sentado ante su mesa en el despacho oval, anuncia que, desgraciadamente, Neil Armstrong y Buzz Aldrin han fallecido tras sufrir un fatal accidente en el Apollo 11. En realidad, este discurso nunca tuvo lugar, sino que fue un *deepfake* concebido por el Instituto de Tecnología de Massachusetts (MIT) (Stockler, 2019) con el fin de alertar sobre los riesgos de la tecnología en la creación de noticias falsas y desinformación, así como para fomentar el espíritu crítico entre el público. Este vídeo del Massachussets Technology Institute (MIT) fue creado utilizando técnicas de aprendizaje profundo y de reemplazo de diálogos en vídeo para recrear el movimiento de los labios de Nixon (Moondisaster, 2019).

El mismo año, Nancy Pelosi protagonizó un *deepfake* en el que aparecía bebida (CBS News, 2019). El vídeo se propagó rápidamente en las redes sociales, si bien, no era técnicamente muy sofisticado.

Un año más tarde, el objetivo de los *deepfakes* fue la reina Isabel II, a la que se hizo dar un curioso mensaje de Navidad (Forrest, 2020) creado por Channel 4 (Channel 4 Entertainment, 2020) para subrayar de nuevo los peligros de la desinformación.

Otro tipo de *deepfake* muy común es el que tiene por objeto a celebridades. Por lo general mujeres, cuyas caras se superponen a otros cuerpos en vídeos pornográficos (Asher Hamilton, Business Insider, 2018) , como el anteriormente citado, protagonizado por Gal Gadot (Jenkins, 2017).

Otro caso de gran repercusión es el de la actriz Scarlett Johansson, habitual víctima de deepfakes de carácter pornográfico (Harwell, 2018). La actriz declaró que intentar protegerse de internet y su perversión era una batalla perdida. También se lamentó de la falta de regulación que permite que cualquiera pueda tomar una imagen, copiarla y pegarla en otro cuerpo, haciéndola parecer totalmente realista gracias a las ayudas digitales, sin tener consecuencias legales. Johansson afirmó que "internet es otro lugar donde el sexo vende y las personas vulnerables son una presa fácil" (Harnell, 2018).

Tom Cruise también protagonizó un *deepfake* famoso (Infobae, 2022), que fue presentado como una parodia, pero era tan realista que resultaba difícil distinguirlo del actor real. Algo similar ocurrió con el *deepfake* en el que Volodymyr Zelensky pedía a las tropas ucranianas la rendición ante el ejército ruso (The Telegragh, 2022). Este *deepfake* señaló el riesgo de manipulación de los discursos de las personas públicas y más en concreto, de los políticos.

También hay que señalar que la detección de *deepfakes* sigue siendo un reto en la actualidad (Seitz-Wald, 2024) como el caso del *deepfake* que mostraba al presidente Joe Biden invitando a los votantes de New Hampshire a no votar en las elecciones primarias o el convincente *deepfake* que llevó a un empleado de la compañía británica Arup a realizar una transferencia de 25 millones de dólares a una cuenta bancaria de Hong-Kong pensando que seguía las órdenes de su jefe (Noto, 2024).

Del mismo modo, en España se pueden destacar algunos *deepfakes* con gran repercusión, como el protagonizado por Pablo Casado, Pedro Sánchez, Pablo Iglesias, Albert Rivera y Santiago Abascal en el que se parodiaba a El Equipo A (Face to Fake, 2019) o el caso Almendralejo (Viejo, 2023) que abrió un debate social sobre la privacidad y la seguridad digital de los menores, ya que se difundieron *deepfakes* protagonizados por chicas menores de edad y creados por un grupo de adolescentes también menores. Otro *deepfake* con gran repercusión, fue el del youtuber Ibai Llanos, promocionando productos que en realidad nunca había mencionado (Errodringer, 2022).

Estos ejemplos muestran cómo los *deepfakes* pueden afectar tanto a figuras públicas como privadas y en diversos ámbitos de la vida, afectando gravemente a la privacidad y la imagen de las personas objeto de los mismos. Igualmente, en los últimos años, el avance de la tecnología de los *deepfakes* ha alcanzado un nivel alarmante de sofisticación, haciendo cada vez más difícil distinguir entre lo real y lo falso. Este progreso plantea serias preocupaciones de seguridad, ya que los *deepfakes* pueden ser utilizados para difundir desinformación, manipular opiniones públicas y cometer fraudes. La capacidad de crear vídeos e imágenes hiperrealistas, que son casi imposibles de detectar a simple vista, pone en riesgo la integridad de la información y la confianza en los medios de comunicación. Además, el potencial uso malintencionado de esta tecnología puede tener consecuencias devastadoras en temas legales, financieros y personales, lo que subraya la necesidad urgente de desarrollar métodos eficaces para su detección y regulación.

6. Tendencias clave, prevalencia e impacto de los deepfakes en el panorama audiovisual actual

En los últimos años, el uso de *deepfakes* ha experimentado un crecimiento significativo, como así demuestran los datos recientes (Markey & Horswell, 2024). Para comprender de forma más profunda el impacto de los *deepfakes*, es crucial revisar algunas estadísticas clave que reflejan su prevalencia. Según el informe llevado a cabo en Estado Unidos por una agencia de seguridad digital, el número total de vídeos *deepfake* presentes en las redes durante el año 2023 fue de 95.820 representando un crecimiento de 550% sobre el año 2019 (Security Hero, 2023). Es relevante señalar que el 98% de los *deepfakes* que

se encuentran online son vídeos de contenido pornográfico explícito. El mismo estudio aporta la cifra de 303.640.207 visualizaciones obtenidas por las diez plataformas más populares de pornografía basada en *deepfakes.*

En otra reciente encuesta realizada por Ashraf (2024), el 74,3% de los estadounidenses encuestados manifestó su preocupación por la posible manipulación de la opinión pública a través de *deepfakes* y un 65,7% tiene la creencia de que *deepfakes* publicados durante las campañas electorales podrían influir decisivamente en la opinión de los votantes. Asimismo, el 48,6% manifestó una disminución significativa en la confianza hacia los vídeos en línea o el contenido de los medios debido a los *deepfakes.*

Los *deepfakes* también inciden de gran manera en otras áreas, como la de la seguridad financiera, sector que se enfrentó a retos sin precedentes, ya que centró los ataques y los fraudes utilizando *deepfakes* creados con IA, aumentando un 3000% en 2023 y alcanzando su máximo histórico en enero de 2024 (Markey & Horswell, 2024). Se constata, por tanto, que los *deepfakes* representan una amenaza particularmente nociva para todas aquellas estrategias de prevención y detección del fraude basadas en comprobaciones biométricas (Onfido, 2025), tales como una foto estática tipo *selfie* o un vídeo para realizar la comprobación de identidad de una persona, a tenor de los porcentajes de intentos de fraudes biométricos que superan el 40%.

Asimismo, en un reciente estudio llevado a cabo entre los *C-suite executives*, es decir, aquellos que ostentan los cargos de CEO (*Chief Executive Officer*), CFO (*Chief Financial Officer*), COO (*Chief Operating Officer*) de las principales empresas de Estados Unidos, muchos de ellos reconocieron el poder destructivo de los *deepfakes* y que el impacto de los daños por ataques exitosos alcanzó incluso hasta el 10% de las ganancias anuales de las empresas (Brooks, 2024). El mismo informe concluye que son pocas las compañías que han articulado acciones para mitigar el riesgo de los *deepfakes* en su negocio, poniendo de manifiesto que el 80% de las empresas no tienen protocolos para manejar ataques de *deepfakes* y más del 50% de los directivos admitió que sus empleados no cuentan con las herramientas ni la capacitación suficiente para reconocer o manejar los ataques de *deepfakes*.

En este mismo sentido, un estudio (Mai et al., 2023) realizado sobre *deepfakes* de audio en los que se imita el discurso humano, reveló que una cuarta parte de los participantes en el estudio no eran capaces de diferenciar las grabaciones reales del audio *deepfake*, lo que enlaza y confirma los resultados de una encuesta llevada a cabo por MacAfee (2023) para medir el impacto de los fraudes generados por *deepfakes* de audio, en el que el 70% de los encuestados afirmó no confiar en su capacidad para distinguir una voz real de una clonada artificialmente, inmediatamente tras lo cual, el 40% aseveró que, sin duda, ayudaría si recibiera un mensaje de audio de su pareja en apuros pidiendo su ayuda. Esta encuesta (Chole et al., 2023) llevada a cabo en nueve países distintos tales como Estados Unidos, Méjico, Brasil, Reino Unido, Alemania, India, Francia, Australia y Japón sobre una muestra de 7000 personas, de las cuales, hasta el 53% dijeron compartir sus voces en línea o a través de notas grabadas, al menos una vez a la semana. Por este motivo, *Youtube*, podcasts, *reels* o redes sociales son fuentes comunes de grabaciones de audios auténticos, que son utilizadas por los ciberdelincuentes para obtener voces de personas reales, que luego son clonadas y utilizadas para enviar mensajes-estafa en busca de dinero. De hecho, una de cada diez personas confirma haber recibido alguna vez mensajes de ese tipo y el 77% lamenta haber sido estafado a causa de haber creído la veracidad del mensaje.

En conclusión, se ha producido un incremento exponencial en los fraudes basados en *deepfakes* entre 2023 y 2024 (Cruz, 2024) con más de 95.000 vídeos de *deepfake* en línea, de los cuales el 98% es de contenido pornográfico explícito. Del mismo modo, los fraudes financieros basados en *deepfakes* alcanzaron su máximo histórico en 2024, tras un aumento del 3000% en el año 2023. Esta puede ser la razón que sustenta la preocupación expresada por el 74,3% de los encuestados (Security Hero, 2023) respecto a la manipulación de la opinión pública por los *deepfakes,* subrayando la urgencia de desarrollar métodos efectivos para detectar y mitigar el impacto de esta tecnología.

7. Deepfakes y su uso malintencionado. Análisis de la amenaza

El desarrollo y la propagación de los *deepfakes* han generado una creciente preocupación debido a su potencial uso malintencionado. Los *deepfakes* se utilizan para una variedad de fines maliciosos, incluyendo campañas de desinformación, chantaje, suplantación de identidad y acoso. Estas tecnologías permiten la creación de contenido audiovisual falsificado con tal grado de realismo, que resulta difícil, incluso para expertos, distinguir entre lo auténtico y lo manipulado. Esta capacidad de falsificación ha sido utilizada en numerosos contextos para llevar a cabo fraudes, difundir desinformación y comprometer la seguridad personal y pública. La posibilidad de generar vídeos e imágenes convincentes (García-Ull, 2021) con objetivos engañosos representa una amenaza significativa en esferas como la política, la economía y la vida privada de los individuos. Por tanto, el estudio y la implementación de estrategias de detección y mitigación de *deepfakes* se torna imperativo para salvaguardar la integridad (Alanazi & Asif, 2024) de la información y proteger a la sociedad de los efectos negativos de esta tecnología.

En línea con lo que muestran los estudios, los principales analistas de tendencias digitales (Juniper Research, 2023); (Sumsub, 2024) coinciden al afirmar en todos sus informes que los fraudes relacionados con la manipulación de audios o vídeos, en particular, los *deepfakes* constituyeron la principal fuente de delitos relacionados con la identidad de las personas. Insisten en que, a medida que la tecnología avanza, el uso de IA para manipular audio y vídeos reales, se crea contenido falso cada vez más convincente (Bañuelos, 2022) y difícil de detectar.

Para comprender el verdadero peligro de los *deepfakes*, es imprescindible una alfabetización adecuada sobre su naturaleza. Según encuestas a nivel mundial, Alemania y España son los países que menos conocen qué es un *deepfake,* ya que el 75%

de los encuestados en estos países admitieron no saber qué es (iProov, 2023).

Aunque el conocimiento sobre los *deepfakes* varía considerablemente de un país a otro, la falta de información sobre el tema subraya la necesidad de campañas educativas y de concienciación específicas en diferentes regiones. Este hallazgo sugiere que una parte significativa de la población es susceptible de caer en estafas de audio *deepfake*, ya que solo el 25% de la población española admitió conocer lo que es. Esto pone de relieve la urgencia de aumentar la educación pública y el desarrollo de herramientas de detección sofisticadas.

Si bien la percepción de la peligrosidad de los *deepfakes* ha sido confirmada por el 62% de los encuestados, para ser consciente de la peligrosidad de los *deepfakes*, es fundamental entender sus posibles usos potenciales. Estos usos son variados e incluyen venganza, con el fin de causar daño reputacional o escarnio público; chantaje con fines económicos; suplantación de identidad para cometer fraudes y estafas, incluido el fraude bancario; creación de pornografía no consentida; campañas de desinformación, noticias falsas e interferencia electoral; y acoso escolar y personal. La imposibilidad de los seres humanos para detectar los *deepfakes* de audio es un aspecto preocupante. Una cuarta parte de los encuestados en un estudio realizado por Mai et al. (2023), afirmó no poder diferenciar una grabación de audio real de un *deepfake*. Esta estadística es alarmante, teniendo en cuenta las posibles aplicaciones del audio *deepfake* para perpetrar estafas y difundir información falsa. Estas preocupantes cifras subrayan la necesidad urgente de implementar medidas de protección y desarrollar métodos efectivos para localizar *deepfakes*. La alfabetización sobre el tema y la creación de herramientas avanzadas de detección son esenciales para disminuir los riesgos (García-Ull, 2021) asociados a esta tecnología emergente.

Las estadísticas presentadas revelan un panorama polifacético de imágenes generadas por IA y *deepfakes*. El rápido aumento de los fraudes relacionados con *deepfakes*, las diferencias globales en el conocimiento del tema y la capacidad variable del público para reconocer estas falsificaciones destacan áreas urgentes de atención. Aunque los avances tecnológicos han democratizado las capacidades de la IA, también amplifican la necesidad de directrices éticas y de concienciación. A medida que se avanza tecnológicamente, equilibrar el potencial innovador de los *deepfakes* con su impacto social requiere esfuerzos concertados en educación, regulación y desarrollo tecnológico. En el mismo sentido, este campo en evolución exige vigilancia continua y un compromiso responsable de todas las partes interesadas en el mundo digital, para mitigar los riesgos y maximizar los beneficios de esta tecnología emergente.

8. La detección de deepfakes: el gran desafío

A tenor de los datos aportados con anterioridad, se pone de manifiesto tanto la gran calidad técnica alcanzada por los *deepfakes,* como la extrema dificultad manifestada por los ciudadanos de distinguir el contenido real del artificial.

En la actualidad se han articulado diversos métodos de identificación de *deepfakes,* con diferentes niveles de eficacia (Westerlund, 2019). Uno de ellos es el análisis de metadatos, esto es, el análisis concienzudo de la información textual sobre la producción del archivo multimedia (tipo de cámara, fecha de creación, ISO, programa de edición utilizado, etc.) con el fin de evaluar si la imagen puede haber sido manipulada. Este método no es del todo eficaz, ya que en ocasiones es complicado identificar las manipulaciones realizadas sobre los archivos. Además, los creadores de *deepfakes* pueden alterar o eliminar metadatos para evitar la detección. Otro problema es que los metadatos pueden ser fácilmente falsificados o manipulados para parecer auténticos. Los *deepfakes* más avanzados pueden incluir metadatos que coincidan perfectamente con los parámetros esperados, haciendo que el análisis de metadatos sea inútil en estos casos. Por último, el análisis de metadatos no puede detectar manipulaciones en el contenido visual o auditivo del archivo, lo que limita su eficacia en la identificación de *deepfakes.* Otro sistema de detección es el de *Error Level Analysis* (ELA), relacionado con la compresión de las imágenes, que sufren distorsiones debido a los algoritmos utilizados en la compresión (Matern et al., 2019). Sin embargo, este método también tiene limitaciones. En primer lugar, los *deepfakes* más sofisticados pueden minimizar las distorsiones visibles, haciendo que el ELA sea menos efectivo. Además, el ELA puede generar falsos positivos, ya que las imágenes auténticas también pueden mostrar niveles de error debido a la compresión natural y a los procesos de edición.

Por último, los creadores de *deepfakes* están constantemente mejorando sus técnicas para evadir estos métodos de detección, lo que hace que la identificación de *deepfakes* sea un reto continuo.

A pesar de ello, se señalan algunos aspectos a tener en cuenta a la hora de identificar *deepfakes* en formato de vídeo (Matern et al., 2019), que pudieran levantar sospechas sobre su autenticidad, como son:

- Errores de geometría en la morfología de las figuras y los rostros humanas presentados, especialmente visibles en los límites faciales.
- Imprecisiones en las formas, que se pueden presentar levemente distorsionadas.
- Falta de coherencia lumínica. Apreciándose diferente luz a lo largo del vídeo.
- Episodios de heterocromía, con casos de diferentes colores de ojos en personas o animales.
- Excesiva perfección general, con carencia de las imperfecciones típicas del mundo real, como ejemplo, las arrugas en la piel.

La identificación de los *deepfakes* de audio, tal y como han revelado las encuestas (Mai K. y otros, 2023) es más complicada aún que la detección de los *deepfake* de vídeo (Johnson, 2020). Aun así, los expertos aportan algunas claves para localizar audios creados artificialmente:

- La duración del audio influye. Cuanto más corto es el audio, más complicado será dilucidar si es auténtico o no.
- La calidad del sonido de fondo puede dar información sobre su veracidad.
- La pronunciación de algunos sonidos puede ser también un indicador de engaño, ya que la producción de algunos sonidos es complicada para las IA.
- Lo mismo ocurre con la rapidez del discurso. A los sistemas de creación de *deepfakes* les resulta difícil imitar velocidades altas. Cuanto más alta sea la velocidad, más complicada la creación artificial del audio.

En los últimos tiempos, la sofisticación en la creación de *deepfakes* realistas se ha desarrollado en paralelo con la implementación en los sistemas de detección (Omar et al., 2023) como *FaceForensics++* que consiguen hasta un 96,90% de efectividad en la detección de *deepfakes* de vídeo.

9. Implicaciones y retos legales de los deepfakes

En primer lugar, una vez definido el concepto de *deepfake*, sus características, proceso de creación, pautas de detección y amenazas que supone, es conveniente analizar de qué modo se relaciona con los derechos de los ciudadanos, ya que, como se ha mostrado, los *deepfakes* son creaciones audiovisuales artificiales con un altísimo grado de realismo, que pueden afectar de forma muy negativa a las personas.

Por su naturaleza, los *deepfakes* afectan principalmente al derecho a la propia imagen, al honor y a la intimidad como recoge el artículo 18 de la norma fundamental española (Constitución Española, 1978), aunque la jurisprudencia todavía es incipiente en lo relativo a delitos relacionados específicamente con *deepfakes*. De hecho, la regulación que típicamente se esgrime es el Reglamento de Inteligencia Artificial de la Unión Europea (RIA) (Reglamento (UE) 2024/1689, 2024).

Dicho Reglamento señala en su apartado de definiciones) a los *deepfakes* como:

> «Ultrasuplantación»: un contenido de imagen, audio o vídeo generado o manipulado por una IA que se asemeja a personas, objetos, lugares, entidades o sucesos reales y que puede inducir a una persona a pensar erróneamente que son auténticos o verídicos (Reglamento (UE) 2024/1689, 2024, art. 3.60)

Legalmente, los *deepfakes*, tal y como vienen definidos en la norma, presentan un gran número de aristas y problemas que pueden incidir y afectar a numerosos aspectos relacionados con los derechos de las personas. De este modo, los *deepfakes* pueden constituir delitos tales como injurias, en tanto en cuanto las injurias son acciones o expresiones que lesionan la dignidad de una persona, menoscabando su fama o atentando contra su propia estimación (Artículo 208 del Código Penal, 1995).

Igualmente, aunque no se trate estrictamente de imágenes reales, sino creadas artificialmente, su difusión sí puede afectar a la intimidad de las personas objeto del *deepfake* (Ley Orgánica 1/1982, 1982). Asimismo, entra dentro de lo posible que el fin último del *deepfake* sea el lucro, con lo que podría entrar dentro del delito de extorsión, tipificado en el artículo 243 del Código Penal, (1995), ampliándose a la vulneración de los artículos 197 y 198, de vulneración de la intimidad del menor y pornografía simulada por IA respectivamente, en el caso de que los afectados por la creación del *deepfake* sean menores.

Adicionalmente, el Reglamento (UE) 2024/1689, (2024) articula un sistema de prevención con el fin de velar por la no vulneración de los derechos antes mencionados, relacionados con el derecho al honor, la intimidad y la propia imagen (Ley Orgánica 1/1982, 1982). Por un lado, el Reglamento (UE) 2024/1689, 2024 (art. 50.4) establece que las personas físicas o jurídicas responsables del despliegue, que usen un sistema de IA bajo su propia autoridad, tienen la obligación de hacer público que esos contenidos han sido generados de forma artificial y deben insertar una "marca de agua" que así lo advierta. Por otro lado, el Reglamento (UE) 2024/1689, 2024 (art. 50.2) establece que los proveedores de sistemas de IA deben velar por que los resultados del sistema de IA estén señalados en formatos legibles por máquinas de tal manera que se pueda detectar claramente si han sido generados o manipulados artificialmente.

En lo que respecta a los afectados o afectadas por los *deepfakes*, deben acudir de inmediato a tramitar una denuncia ante las Fuerzas y Cuerpos de Seguridad del Estado, quienes se encargarán de dar curso a la denuncia y de iniciar las acciones e investigaciones pertinentes. Además, según la naturaleza del *deepfake*, se puede acudir al Canal Prioritario (AEPD, s.f.) de la Agencia Española de Protección de Datos (AEPD), tan pronto como se tenga conocimiento de alguna publicación de vídeos, fotografías o audios de contenido sensible, sexual o violento. Este canal puede solicitar la retirada de dichos contenidos si se estima que su difusión es una vulneración grave de los derechos y libertades de la persona o que conlleva riesgos para su salud física o mental.

Por último, el Instituto Nacional de Ciberseguridad (INCIBE, s.f.) cuenta con servicio de respuesta a incidentes de ciberseguridad, que tiene entre sus funciones proporcionar soporte técnico e información sobre incidencias relacionadas con la ciberseguridad, por lo que es muy recomendable notificar lo sucedido también a este organismo estatal.

Todo lo anteriormente mencionado, tiene por objeto crear un marco legal que facilite el uso responsable de las tecnologías, protegiendo a los individuos de posibles abusos, al mismo tiempo que se garantizan sus derechos fundamentales, persiguiendo el uso ilícito o la difusión de imágenes o voces modificadas, si bien se han identificado posibles riesgos en conexión con la incidencia de los *deepfakes* sobre los derechos al honor, la intimidad y la propia imagen.

10. Conclusiones

La exhaustiva revisión bibliográfica realizada y el análisis detallado de las estadísticas presentadas subrayan la creciente amenaza que representan los *deepfakes* en determinadas áreas de la vida de las personas y la necesidad urgente de desarrollar

tecnologías de detección y marcos legales que den lugar a una regulación adecuada para mitigar sus riesgos y afrontar los desafíos éticos que plantean los *deepfakes*.

La hipótesis central toma como punto de partida de esta revisión bibliográfica estableció que la Inteligencia Artificial (IA) y los *deepfakes* están transformando radicalmente el fenómeno de la información, hipótesis que ha sido confirmada al destacar cómo estas tecnologías permiten la creación rápida y fácil de contenido falso, erosionando la confianza pública en la veracidad de la información y socavando la integridad de las instituciones democráticas. Esto demuestra que las IA y los *deepfakes* están redefiniendo la forma en que se distribuye la información y los desafíos que esto plantea para la sociedad.

En segundo término, se planteó la cuestión del impacto social y ético de los *deepfakes* y hasta qué punto representan una amenaza para la privacidad y la seguridad, especialmente en lo relacionado con la autenticación biométrica. Ante lo cual, se puede concluir que los *deepfakes* plantean desafíos significativos para la privacidad personal y la seguridad de la información, respaldando la hipótesis inicialmente planteada, señalando que resultan especialmente dañinos al saltarse fácilmente los protocolos de seguridad relacionados con la biometría y la identidad digital.

En conexión con lo anteriormente expuesto, la detección de los *deepfakes* se erige como un reto considerable, siendo crucial desarrollar algoritmos más sofisticados y métodos de verificación para mitigar los riesgos asociados con estos contenidos falsos, alineándose con la hipótesis planteada.

Igualmente, se estableció la hipótesis de la urgencia de delimitar marcos legales que protejan la privacidad y la seguridad de las personas ante la facilidad de crear y difundir *deepfakes*. La facilidad con la que se pueden crear y difundir subraya la necesidad urgente de configurar marcos legales. Las conclusiones confirman esta hipótesis, destacando la importancia de adaptar rápidamente las leyes para abordar los nuevos desafíos que presentan estas tecnologías.

La falta de conocimiento sobre los *deepfakes* entre la población general destaca la necesidad de establecer mecanismo de información a los ciudadanos. Las encuestas y estadísticas revelaron un desconocimiento generalizado sobre los *deepfakes,* lo que pone de manifiesto la necesidad de campañas educativas y de concienciación. Los datos respaldan esta hipótesis, enfatizando la importancia de informar al público sobre los riesgos para reducir su impacto negativo.

Se pone de manifiesto la necesidad de abordar los desafíos que representan los *deepfakes* desde múltiples frentes, incluyendo la tecnología, la legislación y la educación, con el objetivo de proteger la integridad de la información en la era digital. El análisis detallado de las estadísticas y la revisión bibliográfica realizados subraya la creciente amenaza que representan los *deepfakes* y la necesidad urgente de desarrollar tecnologías de detección y marcos legales para mitigar sus riesgos.

Igualmente, respecto a la regulación normativa, actualmente España se encuentra en proceso de desarrollo del Reglamento de Inteligencia Artificial (RIA) de la Unión Europea. (Reglamento (UE) 2024/1689, 2024). Este reglamento establece un marco regulador armonizado para la inteligencia artificial en toda la UE, y los Estados miembros, incluida España deben cumplir con estas normas. Sin embargo, queda abierto el debate sobre si esta regulación normativa es suficiente para paliar los daños que se pueden generar con los nuevos contenidos.

11. Limitaciones

La principal limitación observada radica en la misma naturaleza de este artículo, en la medida en que se trata de una revisión bibliográfica, con lo que la falta de datos empíricos podría afectar a la validez de algunas conclusiones.

La rápida evolución tecnológica en el campo de los *deepfakes,* así como el dinamismo de la Inteligencia Artificial podrían significar que algunos de los datos presentados pudieran quedarse obsoletos con rapidez.

En el artículo se plantean los impactos éticos y sociales de los *deepfakes* de manera general. Sin embargo, el fenómeno de los *deepfakes* puede presentar variaciones según los diferentes contextos sociales y culturales.

Se observan limitaciones en las medidas de detección y mitigación propuestas, ya que pueden resultar insuficientes o demasiado generales y la implementación práctica de soluciones requeriría de un análisis más detallado y específico, así como estudios de caso que demostrasen su efectividad en diferentes escenarios.

El tema abordado en este artículo es extremadamente amplio y complejo, por lo que no ha sido posible profundizar en todos los aspectos con el detalle deseado. Algunos puntos tratados merecen un análisis más exhaustivo, que se espera poder abordar en futuras publicaciones.

Estas limitaciones subrayan la necesidad de un enfoque más exhaustivo y actualizado en la investigación sobre *deepfakes*, así como la importancia de considerar múltiples perspectivas y contextos para abordar de manera efectiva los desafíos que presentan estas tecnologías emergentes.

12. Prospectiva

La tecnología detrás de los *deepfakes* y la Inteligencia Artificial seguirá evolucionando, permitiendo la creación de contenidos aún más realistas y difíciles de detectar. Es probable que se produzcan mejoras en la calidad y el realismo de los *deepfakes*, así como en la eficiencia de los algoritmos utilizados para generarlos. Es de esperar que a medida que los *deepfakes* se vuelvan más sofisticados, también lo hagan las herramientas de detección, desarrollando algoritmos más avanzados y técnicas de verificación que puedan identificar características sutiles que delaten las falsificaciones.

La necesidad de marcos legales robustos para abordar los desafíos que presentan los *deepfakes* será cada vez más urgente. Es probable que los gobiernos y las organizaciones internacionales trabajen en conjunto para establecer regulaciones que protejan la privacidad y la seguridad de las personas.

La educación y la concienciación pública sobre los *deepfakes* serán fundamentales para mitigar su impacto negativo. Se espera que se implementen campañas educativas a nivel global para informar al público sobre los riesgos y las señales de los *deepfakes.*

A pesar de los riesgos, los *deepfakes* también tienen aplicaciones positivas, como en la industria del entretenimiento y la educación. Es probable que en un futuro se produzca un aumento en el uso ético de esta tecnología para crear contenidos innovadores y educativos.

En lo que respecta a líneas futuras de investigación, una de ellas se podría centrar en desarrollar algoritmos más precisos y eficientes para detectar *deepfakes.* Esto incluirá el uso de técnicas de aprendizaje automático y redes neuronales para identificar patrones y anomalías en los contenidos generados.

Es crucial investigar el impacto social y psicológico de los *deepfakes* en la población. Esto debería incluir estudios sobre cómo afectan a la percepción pública, a la confianza en los medios y las instituciones, y al bienestar emocional de las personas.

Una investigación en el ámbito legal y normativo será esencial para establecer marcos que regulen el uso de los *deepfakes*. Esto incluirá el análisis de las implicaciones legales y la creación de políticas que protejan a los individuos y las organizaciones.

En el ámbito concreto de este artículo resultaría de gran interés fomentar la interdisciplinariedad en la investigación: con la colaboración entre diferentes disciplinas, como la informática, la psicología, el derecho y las ciencias sociales, para abordar de manera integral los desafíos que presentan los *deepfakes.* Una investigación interdisciplinaria permitirá desarrollar soluciones más completas y efectivas. Estas prospectivas y líneas de investigación subrayan la importancia de abordar los desafíos y oportunidades que presentan los *deepfakes* desde múltiples perspectivas, asegurando un enfoque equilibrado y responsable en el uso de esta tecnología emergente.

Tal y como se indicó con anterioridad, las implicaciones de los *deepfakes* pueden diferir significativamente según la región y la cultura, lo que no se explora en profundidad. También deberían abordarse adecuadamente las variaciones en diferentes contextos culturales y sociales.

13. Referencias

2023 State of Deepfakes. Realities, Threats and Impact. (2023). *Security Hero*. Estados Unidos. https://www.securityhero.io: https://www.securityhero.io/state-of-deepfakes/#key-findings

AEPD. (s.f.). *Agencia Española de Protección de Datos*. Retrieved diciembre de 2024, from AEPD Canal Prioritario: https://www.aepd.es/canalprioritario

Ajder, H., Cavalli, F., Patrini, G., & Cullen, L. (2019). The state of Deepfakes: Landscape, threats, and impact.

Alanazi, S., & Asif, S. (18 de septiembre de 2024). Exploring deepfake technology: creation, consequences and countermeasures. *Human-intelligent systems integration, 6*, 49-60. https://doi.org/10.1007/s42454-024-00054-8

American Cinematographer. (noviembre de 1973). *11*(54).

Asher Hamilton, I. (3 de diciembre de 2018). *Business Insider*. https://www.businessinsider.com/scarlett-johansson-stopping-deepfake-porn-of-me-is-a-lost-cause-2018-12

Asher Hamilton, I. (3 de diciembre de 2018). *Business Insider*. Retrieved noviembre de 2024, from Scarlett Johansson says trying to stop people making deepfake porn videos of her is a 'lost cause'

Ashraf, S. J. (2024). Tendencias y amenazas de deepfake. VPNRanks. https://www.vpnranks.com/es-es/recursos/tendencias-y-amenazas-deepfake/

Bañuelos, J. (2022). Evolución del Deepfake: campos semánticos. *ICONO 14 Revista de comunicación y tecnologías emergentes*(20). https://doi.org/https://doi.org/10.7195/ri14.v20i1.1773

Batchen, G. (1999). *Burning with Desire: The Conception of Photography.* MIT Press.

Brooks, C. (28 de junio de 2024). 1 in 10 Executives Say Their Companies Have Already Faced Deepfake Threats. *Business.com*. https://www.business.com/articles/deepfake-threats-study/

Brugioni, D. (1999). *Photo fakery: The history and techniques of photographic deception and manipulation.* Brassey's.

CBS News. (25 de Mayo de 2019). *CBS News.* https://www.cbsnews.com/news/doctored-nancy-pelosi-video-highlights-threat-of-deepfake-tech-2019-05-25/

Channel 4 Entertainment. (25 de diciembre de 2020). *Channel 4.* Retrieved diciembre de 2024, from Youtube.com: https://www.youtube.com/watch?v=IvY-Abd2FfM

Chole, V., Crimmins, C., DeVane, O., & Karnik, A. (2023). Beware the artificial impostor: A McAfee cybersecurity artificial intelligence report. McAfee. https://www.mcafee.com/content/dam/consumer/en-us/resources/cybersecurity/artificial-intelligence/rp-beware-the-artificial-impostor-report.pdf

Coates, J. (2018). *Photographing the Invisible; Practical Studies in Spirit Photography, Spirit Portraiture, and Other Rare But Allied Phenomena.* Creative Media Partners, LLC.

Código Penal, Artículo 208. (1995). Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal. https://www.boe.es/buscar/doc.php?id=BOE-A-1995-25444

Código Penal, Artículos 189 y 243. (1995). Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal. https://www.boe.es/buscar/doc.php?id=BOE-A-1995-25444

Consejo de Europa. (18 de febrero de 2019). https://www.consilium.europa.eu/es/press/press-releases/2019/02/18/european-coordinated-plan-on-artificial-intelligence/pdf

Constitución Española. (1978). *BOE.* https://www.boe.es/buscar/doc.php?id=BOE-A-1978-31229

Crichton, M. (Dirección). (1973). *Westworld* [Película].
Cruz, B. (26 de septiembre de 2024). *https://www.security.org.* https://www.security.org/resources/deepfake-statistics/#what
Cruz, B. (26 de septiembre de 2024). *https://www.security.org.* Retrieved octubre de 2024, from https://www.security.org/resources/deepfake-statistics/
Dagar, D., & Vishwakarma, D. (2022). A literature review and perspectives in deepfakes: generation, detection, and applications. *International Journal of Multimedia Information Retrieval, 11,* pages 219–289. https://doi.org/https://doi.org/10.1007/s13735-022-00241-w
Damer , N., Saladié , A., & Kuijper, A. (2018). Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network. *EEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS).*
Entrust. (2025). *Onfido.* https://onfido.com: https://onfido.com/landing/identity-fraud-report/
EPRS. European Parliamentary Research Service . (julio de 2021). *https://www.europarl.europa.eu.* Retrieved diciembre de 2024, from https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf
Errodringer . (29 de julio de 2022). Retrieved diciembre de 2024, from Youtube: https://www.youtube.com/watch?v=N6KeltkxXNs
Face to Fake. (11 de noviembre de 2019). *Face to Fake.* Youtube.com: https://www.youtube.com/watch?v=dj5M4s-cdAw
Fineman, M. (2012). *Faking it : manipulated photography before Photoshop.* New York: Metropolitan Museum of Art.
Forrest, A. (24 de diciembre de 2020). *Independent.* https://www.independent.co.uk/news/uk/home-news/queen-deepfake-channel-4-christmas-message-b1778542.html
García-Ull, F. (junio de 2021). Deepfakes: el próximo reto en la detección de noticias falsas. *Anàlisi: Quaderns de Comunicació i Cultura*(64), 103-120. https://doi.org/https://doi.org/10.5565/rev/analisi.3378
Goodfellow, I. P.-A.-F. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*(27), 2672-2680.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., & Ozair, S. (2014). Generative Adversarial Nets. (U. d. Montreal, Ed.) *https://journals.plos.org.* https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
Gottfried, J. (14 de junio de 2019). *Pew Research Center.* https://www.pewresearch.org/short-reads/2019/06/14/about-three-quarters-of-americans-favor-steps-to-restrict-altered-videos-and-images/
Harwell, D. (30 de diciembre de 2018). Fake-porn videos are being weaponized to harass and humiliate women: 'Everybody is a potential target'. *The Washington Post.* https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/
INCIBE. (s.f.). *INCIBE CERT.* Retrieved diciembre de 2024, from https://www.incibe.es/incibe-cert/incidentes/respuesta-incidentes
Infobae. (diciembre de 2022). *https://www.infobae.com/america/tecno/.* Infobae America Tecno: https://www.infobae.com/america/tecno/2022/12/09/los-4-casos-famosos-de-suplantacion-de-identidad-o-deepfake-tom-cruise-y-zuckerberg-figuraron/
iProov. (26 de agosto de 2023). *iProov.* Retrieved noviembre de 2024, from iProov: https://www.iproov.com/blog/deepfakes-statistics-solutions-biometric-protection
Jenkins, P. (Dirección). (2017). *Wonder Woman* [Película].
Johnson, D. (3 de agosto de 2020). Audio Deepfakes: Can Anyone Tell If They're Fake? *How to geek*. https://www.howtogeek.com/682865/audio-deepfakes-can-anyone-tell-if-they-are-fake/
Juniper Research. (2 de diciembre de 2023). *https://www.juniperresearch.com.* https://www.juniperresearch.com/research/fintech-payments/identity/digital-identity-verification-research-report/
Kaplan, L. (2005). *American Exposures: Photography and Community in the Twentieth Century.* University of Minnesota Press.
Kaplan, L. (2008). *The strange case of William Mumler spirit photographer.* University of Minnesota Press.
Karras, T. A. (2019). Style-Based Generator Architecture for Generative Adversarial Networks. . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2947-2958.
Leeder, M. (2015). *Cinematic Ghosts.Haunting and Spectrality from Silent Cinema to the Digital Era.* Bloomsbury Publishing.
Ley Orgánica 1/1982, de 5 de mayo, de protección civil del derecho al honor, a la intimidad personal y familiar y a la propia imagen. (1982). Boletín Oficial del Estado, núm. 115. https://www.boe.es/buscar/act.php?id=BOE-A-1982-11196
MacAfee. (Mayo de 2023). *Artificial Intelligence Survey Reports.* https://www.mcafee.com: https://www.mcafee.com/content/dam/consumer/en-us/resources/cybersecurity/artificial-intelligence/rp-beware-the-artificial-impostor-report.pdf
Mai, K., Bray, S., Davies, T., & Griffin, L. (2023). Warning: Humans cannot reliably detect speech deepfakes. (U. T. ogan Jaya Kumar, Ed.) *PLOS ONE, 18.* https://doi.org/https://doi.org/10.1371/journal.pone.0285333
Mai, K., Bray, S., Davies, T., & Griffin, L. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLoS ONE, 18(8).* https://doi.org/https://doi.org/10.1371/journal.pone.0285333

Mallet, J., Pryor, L., Dave, R., & Vanamala, M. (2023). Deepfake detection analyzing hybrid dataset utilizing CNN and SVM. arXiv. https://doi.org/10.48550/arXiv.2302.10280

Markey, J., & Horswell, S. (22 de Noviembre de 2024). https://www.entrust.com/blog/2024/11/rise-of-sophisticated-fraud-and-deepfakes-at-scale

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. *IEEE Winter Applications of Computer Vision Workshops (WACVW).* https://doi.org/10.1109/WACVW.2019.00020

Méliès, G. (Dirección). (1902). *A trip to the moon* [Película].

Moondisaster. (22 de noviembre de 2019). In the event of moon disaster: https://moondisaster.org/

Noto, G. (17 de mayo de 2024). *CFODive*. https://www.cfodive.com/news/scammers-siphon-25m-engineering-firm-arup-deepfake-cfo-ai/716501/

Omar, K., Sakr, R., & Alrahmawy, M. (2023). An ensemble of CNNs with self-attention mechanism for DeepFake video detection. *Neural Computing and Applications, 36*, 2749-2765. https://doi.org/10.1007/s00521-023-09196-3

Ostendorf, L. (1998). *incoln's Photographs: A Complete Album.* Dayton, OH: Rockywood Press.

Proposición de Ley Orgánica de regulación de las simulaciones de imágenes y voces de personas generadas por medio de inteligencia artificial. (13 de octubre de 2023). *BOE Boletín Oficial de las Cortes Generales(23-1)*. Retrieved diciembre de 2024, from https://www.congreso.es/public_oficiales/L15/CONG/BOCG/B/BOCG-15-B-23-1.PDF

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks.

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican varios reglamentos y directivas. Diario Oficial de la Unión Europea, L 1689, 12 de julio de 2024. HYPERLINK "https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=OJ:L_202401689" https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=OJ:L_202401689

Rosenblum, N. (2008). *A World History of Photography.* Abbeville Press.

Rössler, A., Cozzolino, D., Verdeoliva, L., Riess, C., Thies, J., & Niessner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces. *ArXiv.* https://doi.org/https://doi.org/10.48550/arXiv.1803.09179

Roy, R., Joshi, I., Das, A., & Dantcheva, A. (2022). 3D CNN architectures and attention mechanisms for deepfake detection. HAL. https://hal.science/hal-03524639. https://doi.org/10.1007/978-3-030-87664-7_10

Security Hero. (2023). Retrieved noviembre de 2024, from Security Hero: https://www.securityhero.io/state-of-deepfakes/

Seitz-Wald, A. (24 de febrero de 2024). NBC News: https://www.nbcnews.com/politics/2024-election/biden-robocall-new-hampshire-strategist-rcna139760?_ga=2.181210351.976717714.1719011557-176973521.1719011550

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. En MLResearchPress (Ed.), *32nd International Conference on Machine Learning, PMLR 37:2256-2265, 37.* Retrieved octubre de 2024, from https://proceedings.mlr.press/v37/sohl-dickstein15.html

Spielberg, S. (Dirección). (1993). *Jurassic Park* [Película].

Stockler, A. (03 de diciembre de 2019). MIT Deepfake Video 'Nixon Announcing Apollo 11 Disaster' Shows the Power of Disinformation. *Newsweek*. https://www.newsweek.com/richard-nixon-deepfake-apollo-disinformation-mit-1475340

Sumsub. (2024). *https://sumsub.com.* https://sumsub.com/fraud-report-2024/

Thanh Thi Nguyen, Q. V.-T.-V., & Nguyen, T. (2022). Deep Learning for Deepfakes Creation and Detection: A Survey. *Computer Vision and Pattern Recognition.*

The Telegragh. (17 de marzo de 2022). *The Telegraph*. Youtube: https://www.youtube.com/watch?v=X17yrEV5sl4&t=5s

Unión Europea. (17 de febrero de 2022). *http://data.europa.eu/eli/reg/2022/2065/oj.* (C. d. Europa, Ed.) https://doi.org/DOUE-L-2022-81573

Viejo, M. (23 de octubre de 2023). El caso de los desnudos con IA de Almendralejo se dispara: 26 menores implicados y 21 chicas afectadas. *El País*. https://elpais.com/noticias/almendralejo/

VPNRanks. (2024). *https://www.vpnranks.com.* Retrieved 2024, from https://www.vpnranks.com/es-es/recursos

Washington Post. (31 de diciembre de 2018). *Washington Post*. Retrieved noviembre de 2024, from https://www.washingtonpost.com/technology/2018/12/31/scarlett-johansson-fake-ai-generated-sex-videos-nothing-can-stop-someone-cutting-pasting-my-image/

Westerlund, M. (2019). The emergence of deepfake technology: a review. *Technology Innovation Management Review, 9*(11).

Zhang, T. (enero de 2022). Deepfake generation and detection, a survey. Multimedia Tools and Applications. *Volume 81*, pages 6259–6276. https://doi.org/https://doi.org/10.1007/s11042-021-11733-y