Estafas con deepfakes

Análisis, prevención y casos reales

Índice

Control de versiones

VERSIÓN	FECHA	DESCRIPCIÓN
1.0	02/06/2025	Publicación inicial del informe.

Introducción

En la era digital, el avance vertiginoso de la inteligencia artificial ha dado lugar a tecnologías innovadoras que ofrecen múltiples beneficios, pero también plantean nuevos riesgos y desafíos. Entre estas tecnologías, los deepfakes —videos, audios o imágenes manipulados mediante algoritmos de aprendizaje automático para parecer auténticos— se han convertido en una poderosa herramienta para la creación de fraudes y estafas muy sofisticadas. Esta capacidad para generar falsificaciones hiperrealistas ha abierto la puerta a una nueva generación de delitos digitales que afectan desde la privacidad y la reputación hasta la seguridad financiera y corporativa.

En este informe exploraremos cómo esta tecnología está siendo utilizada para engañar a personas y organizaciones, con especial énfasis en modalidades de fraude como la suplantación de identidad, el fraude financiero y las campañas de desinformación. Además, se analizan casos reales que ilustran la magnitud y el impacto de estas amenazas, así como las estrategias y herramientas disponibles para su detección y prevención.

A medida que los deepfakes se vuelven más accesibles y fáciles de crear, la necesidad de comprender sus riesgos y desarrollar respuestas efectivas se vuelve imperativa para proteger a la sociedad y fortalecer la confianza en la información digital. Este informe busca aportar claridad y conocimiento para enfrentar este fenómeno emergente con rigor y responsabilidad.

¿Qué son los deepfakes?

Los deepfakes son archivos de video, imagen o audio generados mediante inteligencia artificial, diseñados para imitar la apariencia, voz o gestos de una persona con tal realismo que pueden engañar tanto a seres humanos como a sistemas automatizados. El término proviene de la combinación de “deep learning” (aprendizaje profundo) y “fake” (falso), y hace referencia a la utilización de algoritmos avanzados de aprendizaje automático, como las redes generativas antagónicas (GANs), que permiten crear contenidos digitales falsificados de manera convincente.

Esta tecnología funciona alimentando modelos de inteligencia artificial con grandes volúmenes de datos —imágenes, videos o audios reales— para que puedan aprender los patrones y características distintivas de una persona o situación. El resultado son videos, imágenes o audios sintéticos capaces de mostrar a alguien diciendo o haciendo cosas que nunca ocurrieron en la realidad, lo que plantea serios desafíos para la verificación de la autenticidad del contenido digital.

En la actualidad, los deepfakes se han popularizado por su facilidad de acceso y por la variedad de aplicaciones, que van desde el entretenimiento y la publicidad hasta usos maliciosos como la desinformación, la suplantación de identidad y el fraude digital. Su creciente sofisticación y realismo hacen que la detección y prevención de estos contenidos falsos sea cada vez más compleja y relevante para la sociedad.

Tipos de deepfakes: imagen, video y audio

Los deepfakes se presentan en distintas modalidades según el tipo de contenido manipulado: imagen, video y audio. En el caso de las imágenes, la inteligencia artificial puede generar fotografías completamente ficticias desde cero o modificar imágenes existentes para sustituir el rostro de una persona por el de otra.

En el ámbito del video, los deepfakes combinan la manipulación de imágenes con la animación de movimientos faciales y corporales, generando secuencias en las que una persona parece hablar, gesticular o realizar acciones que nunca ocurrieron realmente. Estos videos suelen centrarse en primeros planos para facilitar la edición y suelen tener una duración corta debido a la complejidad técnica que implica su creación.

Por otro lado, los deepfakes de audio, conocidos como deepvoices, utilizan inteligencia artificial para clonar la voz de una persona y hacerle decir frases que jamás pronunció, logrando resultados difíciles de distinguir a simple oído. Esta modalidad se utiliza tanto para suplantación de identidad como para fraudes telefónicos y manipulación de declaraciones públicas.

Principales técnicas utilizadas

Las técnicas empleadas en la creación de deepfakes han evolucionado rápidamente, facilitando tanto la sofisticación de los resultados como el acceso a herramientas que automatizan gran parte del proceso, lo que ha contribuido a su proliferación en redes sociales y entornos digitales.
Las siguientes técnicas, combinadas o utilizadas de forma individual, permiten la creación de deepfakes cada vez más convincentes y difíciles de detectar, representando un desafío creciente para la veracidad de la información digital y la seguridad de las personas.

Redes Generativas Antagónicas (GANs)

Las Redes Generativas Antagónicas (GANs) son la base técnica de la mayoría de los deepfakes. Funcionan mediante la competencia entre dos redes neuronales: una genera contenido falso (imágenes, videos o audios) y la otra evalúa si ese contenido es real o simulado, mejorando ambas con cada iteración. Por ejemplo, las GANs se han utilizado para rejuvenecer digitalmente a actores en películas, como en “The Irishman”, donde los rostros de los protagonistas fueron modificados para parecer más jóvenes.

Lip syncing

La técnica de lip syncing permite modificar los movimientos de los labios de una persona en un video para que coincidan perfectamente con un audio específico, aunque nunca haya pronunciado esas palabras. Esta técnica es muy utilizada en la creación de videos donde se necesita doblar el idioma de los actores manteniendo la sincronía labial, como en películas o anuncios internacionales, o para crear videos en los que una figura pública parece decir algo que nunca dijo.

Face reenactment (recreación facial)

La recreación facial consiste en transferir expresiones y gestos de una persona fuente a otra persona objetivo en un video, logrando que el rostro objetivo imite fielmente los movimientos faciales del original. Esta técnica se ha empleado en aplicaciones de videollamadas y entretenimiento, permitiendo que los usuarios animen avatares digitales en tiempo real o que una persona aparezca realizando gestos o expresiones que nunca realizó.

Voice cloning (clonación de voz)

La clonación de voz utiliza inteligencia artificial para analizar grabaciones de una persona y reproducir su tono, timbre y cadencia, generando audios sintéticos en los que la voz clonada puede pronunciar cualquier texto. Esta tecnología se ha utilizado en fraudes telefónicos donde se suplanta la identidad de directivos para autorizar transferencias bancarias.

Panorama actual de los fraudes con deepfakes

Evolución y tendencias recientes

El fraude con deepfakes ha experimentado un crecimiento explosivo en los últimos años, impulsado por la accesibilidad de herramientas de inteligencia artificial y la sofisticación de las técnicas empleadas. Entre 2022 y 2025, los intentos de fraude con deepfakes han aumentado más de un 2.000%, con un salto especialmente significativo en 2025, donde solo en el primer trimestre se han registrado más incidentes que en todo el año anterior. Este auge no solo responde a la mejora tecnológica, sino también a la proliferación de servicios de "deepfake-as-a-service", que permiten a delincuentes sin conocimientos técnicos ejecutar ataques a gran escala. Además, los deepfakes ya representan entre el 6,5% y el 7% de todos los intentos de fraude digital detectados, y más del 50% de los fraudes con IA involucran la utilización de deepfakes. Las técnicas tradicionales de detección, como el análisis de movimientos o gestos faciales, han quedado obsoletas ante la capacidad de los deepfakes para simular videos y audios en tiempo real, lo que exige nuevas soluciones de ciberseguridad y biometría.

Ámbitos más afectados: empresas, finanzas, política y salud

El sector financiero es uno de los más afectados, con un crecimiento del 2.137% en los intentos de fraude por deepfake en los últimos tres años. Los bancos y servicios de pago se enfrentan a apropiaciones de cuentas, fraudes en pagos con tarjeta y suplantación de identidad mediante videos y audios sintéticos, lo que ha obligado a reconsiderar las estrategias de seguridad y a implementar nuevas tecnologías de detección. En el ámbito empresarial, los ataques de deepfake se han utilizado para engañar a empleados y directivos, logrando transferencias de fondos o acceso a información confidencial a través de videollamadas o mensajes de voz falsificados. La política también ha sido blanco de deepfakes, con casos de manipulación de discursos y campañas de desinformación, como el uso de audios falsos en procesos electorales en Estados Unidos. En el sector salud, aunque menos frecuente, se han detectado intentos de suplantación de identidad para acceder a datos médicos o manipular resultados de pruebas, aprovechando la confianza en sistemas de verificación biométrica.

Distribución geográfica de los casos

La distribución de los fraudes con deepfakes muestra una concentración en regiones con alta penetración digital y acceso a tecnologías avanzadas de inteligencia artificial. Estados Unidos, Europa y China lideran el número de incidentes reportados, tanto en la ejecución de fraudes como en el desarrollo de sistemas de detección y prevención. Sin embargo, el aumento más pronunciado en los últimos años se ha registrado en países como Filipinas y Vietnam, donde los intentos de fraude con deepfakes crecieron un 4.500% y 3.900% respectivamente entre 2022

y 2023. Este fenómeno se está expandiendo rápidamente a otras regiones, impulsado por la democratización de las herramientas de IA y la falta de regulaciones específicas, lo que convierte al fraude con deepfakes en una amenaza global y en constante evolución.

Modalidades de estafas y fraudes con deepfakes

Suplantación de identidad en procesos de verificación

Los deepfakes han revolucionado la suplantación de identidad, especialmente en los procesos de verificación digital y autenticación biométrica. Los ciberdelincuentes utilizan videos y audios sintéticos para eludir controles de reconocimiento facial o de voz, logrando acceder a cuentas bancarias, sistemas corporativos o plataformas de servicios. Por ejemplo, se han documentado casos en los que los atacantes crean identidades sintéticas para solicitar préstamos o abrir cuentas bancarias fraudulentas, o bien emplean deepfakes para apropiarse de cuentas existentes, superando las barreras de seguridad tradicionales. Gartner estimó que en 2023 los deepfakes estuvieron implicados en el 20% de los ataques exitosos de apropiación de cuentas.

Fraudes financieros y bancarios

El sector financiero es uno de los más afectados por los fraudes con deepfakes. Los delincuentes emplean audios o videos falsificados para hacerse pasar por directivos y autorizar transferencias o transacciones fraudulentas. Además, los estafadores utilizan deepfakes para manipular precios de acciones, crear identidades falsas para solicitar productos financieros y realizar fraudes en pagos electrónicos. Un caso emblemático ocurrió en Hong Kong, donde un trabajador fue engañado mediante una videollamada deepfake para transferir 25 millones de dólares, creyendo que hablaba con su director financiero y compañeros de empresa.

Manipulación de informes y comunicaciones corporativas

Las empresas también enfrentan la amenaza de deepfakes en la manipulación de informes y comunicaciones internas o externas. Los atacantes pueden generar videos o audios de ejecutivos anunciando decisiones estratégicas falsas, lo que puede afectar la reputación de la compañía o influir en el valor de sus acciones. En otros casos, se han utilizado deepfakes para suplantar a responsables de recursos humanos en procesos de selección, facilitando el acceso a información sensible o la contratación fraudulenta de personal.

Desinformación y manipulación social

La desinformación es uno de los usos más peligrosos de los deepfakes. Esta tecnología se ha utilizado para crear campañas de manipulación política, como la difusión de audios falsos de líderes mundiales o la generación de imágenes y videos que buscan influir en la opinión pública durante procesos electorales. Ejemplos recientes incluyen audios deepfake de presidentes con declaraciones controvertidas y videos falsos de candidatos en campañas electorales, así como imágenes sintéticas de figuras públicas en situaciones comprometidas para desacreditar su imagen.

Otros usos maliciosos (extorsión, chantaje, etc.)

Además de los fraudes y la desinformación, los deepfakes se utilizan en extorsión, chantaje y acoso digital. Los ciberdelincuentes pueden crear videos o imágenes comprometedoras de víctimas, muchas veces de carácter pornográfico, para exigir dinero o favores a cambio de no difundir el material. También se han documentado casos de espionaje industrial, donde los atacantes emplean deepfakes para obtener información confidencial de empresas, o para explotar a empleados mediante la creación de contenido falso no autorizado.

Casos reales y análisis de incidentes

Uno de los casos más notorios en Europa fue el de una jubilada francesa que, en 2023, fue víctima de una estafa sentimental con deepfakes. Los delincuentes suplantaron la identidad del actor Brad Pitt mediante videos y videollamadas generados por inteligencia artificial, convenciéndola de que mantenía una relación con el famoso. A lo largo de varios meses, la víctima transfirió 830.000 euros para supuestos gastos legales y promesas de un encuentro personal, hasta que finalmente descubrió el engaño.

En el ámbito de la inversión y las criptomonedas, los deepfakes han sido utilizados para crear videos falsos de figuras reconocidas, como Elon Musk. Estos videos, difundidos en redes sociales, promueven inversiones fraudulentas y sorteos inexistentes. Un caso destacado fue el de un jubilado estadounidense que perdió 690.000 dólares tras confiar en un video deepfake de Musk que lo animaba a invertir en un proyecto inexistente.

En el terreno político, la manipulación con deepfakes también ha tenido impacto. En enero de 2024, un audio falso del presidente Joe Biden circuló durante las primarias de New Hampshire, instando a los demócratas a no votar. Este incidente evidenció el potencial de la tecnología para influir en procesos electorales y generar desinformación a gran escala.

Por último, en el sector corporativo, Ferrari logró detectar a tiempo un intento de fraude cuando un ejecutivo recibió mensajes y una llamada deepfake del supuesto CEO Benedetto Vigna. El engaño fue descubierto gracias a una pregunta clave que el impostor no pudo responder, evitando así un posible perjuicio económico y reputacional para la compañía.

Análisis de modus operandi

El modus operandi de los estafadores ha evolucionado con el uso de deepfakes. Los delincuentes recopilan información y material audiovisual de sus objetivos a través de redes sociales o filtraciones de datos, y luego generan contenido sintético convincente. En el fraude del CEO, por ejemplo, se realizan llamadas o videoconferencias en las que el impostor, utilizando un deepfake, solicita información confidencial o autoriza transferencias financieras. En otros casos, se crean identidades sintéticas para superar procesos de verificación

biométrica, abriendo cuentas o accediendo a servicios restringidos. La facilidad de acceso a herramientas de deepfake y la falta de concienciación en la población potencian la efectividad de estos ataques.

Impacto económico y reputacional

El impacto de los fraudes con deepfakes es considerable tanto en términos económicos como reputacionales. El coste medio de un ataque exitoso supera los 400.000 dólares por empresa, y puede ascender a más de 600.000 dólares en el sector financiero, considerando tanto las pérdidas directas como los costes de remediación y cumplimiento normativo.

El daño reputacional es igualmente grave: la manipulación de imágenes, videos o audios puede perjudicar gravemente la imagen pública de empresas y figuras relevantes, y aunque el fraude sea desmentido, el daño a la confianza y la credibilidad ya está hecho. Además, el daño moral y psicológico a las víctimas, especialmente en casos de extorsión o chantaje, representa una proporción significativa de los efectos negativos causados por esta tecnología.

Herramientas y técnicas de detección

Métodos tradicionales y sus limitaciones

Los métodos tradicionales para detectar deepfakes se basan en la observación manual de inconsistencias visuales y auditivas, como movimientos faciales poco naturales, parpadeo irregular, bordes borrosos, cambios de color anómalos o desincronización entre audio y labios. También se emplean procedimientos manuales de verificación de identidad y validación de documentos físicos. Sin embargo, estas técnicas han perdido eficacia frente a la creciente sofisticación de los deepfakes actuales, que logran simular microexpresiones, texturas y movimientos con gran realismo. Además, la revisión manual es lenta, subjetiva y difícil de escalar ante grandes volúmenes de contenido digital, lo que limita su utilidad en entornos empresariales y de alto riesgo.

Soluciones tecnológicas actuales y emergentes

La detección de deepfakes ha avanzado significativamente en los últimos años, impulsada principalmente por el desarrollo de herramientas basadas en inteligencia artificial que analizan imágenes, videos y audios para identificar manipulaciones de manera automatizada y en tiempo real. Estas soluciones emplean técnicas como el análisis de inconsistencias en los patrones faciales, anomalías en el parpadeo, irregularidades en la sincronización labial, y características técnicas propias de los archivos digitales que pueden revelar alteraciones. Además, se están explorando enfoques emergentes que combinan aprendizaje profundo con métodos forenses digitales para mejorar la precisión y rapidez en la identificación de contenidos falsificados.

No obstante, a medida que la inteligencia artificial continúa evolucionando, los deepfakes se vuelven cada vez más sofisticados, realistas y difíciles de distinguir del material auténtico. Esta mejora constante en la calidad de las falsificaciones supera en muchos casos la capacidad de las herramientas de detección actuales, que no siempre son lo suficientemente fiables o precisas para garantizar una identificación definitiva. Por ello, confiar exclusivamente en la tecnología para combatir los deepfakes resulta insuficiente. El sector está apostando por enfoques complementarios, como el uso de hardware especializado para pruebas de vida, la verificación biométrica avanzada y el desarrollo de estándares internacionales de autenticidad, además de la integración de soluciones de detección en dispositivos móviles y plataformas digitales. El mercado de tecnologías de detección de deepfakes continúa expandiéndose rápidamente, impulsado por la necesidad de proteger tanto a usuarios individuales como a empresas frente a un fenómeno en constante evolución.

Prevención y recomendaciones

Protocolos y buenas prácticas de verificación

Las soluciones tecnológicas para la detección de deepfakes, aunque avanzan rápidamente, todavía presentan limitaciones importantes y no son lo suficientemente maduras como para depender exclusivamente de ellas. Los algoritmos pueden ser burlados por nuevas técnicas de manipulación, la calidad de los deepfakes mejora constantemente y los sistemas de detección pueden arrojar falsos positivos o no detectar manipulaciones muy sofisticadas. Por ello, es imprescindible complementar la tecnología con estrategias de protección adicionales, como la autenticación biométrica multimodal, la aplicación de pruebas de vida dinámicas y la validación cruzada de documentos con fuentes oficiales. La intervención de expertos forenses digitales puede ser necesaria en casos críticos, ya que la detección manual sigue siendo relevante ante la sofisticación de los ataques.

Estrategias para empresas

Para evitar fraudes como el del CEO, las empresas deben establecer protocolos internos robustos. Entre las medidas más efectivas se encuentra la verificación de cualquier solicitud sensible —como transferencias de fondos o cambios críticos— a través de un proceso de doble validación: toda instrucción debe ser confirmada por al menos dos empleados de diferentes áreas, utilizando canales de comunicación independientes, por ejemplo, combinando correo electrónico y llamada telefónica directa. Además, es recomendable evitar tomar decisiones importantes basadas únicamente en mensajes de voz, videollamadas o correos electrónicos, sobre todo si provienen de altos directivos. La creación de listas blancas de contactos autorizados, la implementación de políticas de escalamiento y la documentación detallada de cada paso en los procesos críticos refuerzan la protección ante intentos de suplantación.

Estrategias para usuarios

Las personas deben extremar la precaución ante solicitudes o mensajes inesperados, especialmente si involucran datos personales, transferencias de dinero o información confidencial. Es recomendable verificar la identidad del interlocutor por un canal alternativo antes de actuar, y desconfiar de mensajes alarmistas o que requieran acciones urgentes. Es recomendable limitar la cantidad de información personal y multimedia que se comparte en redes sociales y plataformas digitales, ya que reduce el riesgo de que estos datos sean utilizados para crear deepfakes personalizados. Ante la duda sobre la autenticidad de un contenido, es preferible consultar con un experto o reportar el caso a la empresa o autoridad correspondiente.

Capacitación y concienciación

La formación continua es fundamental para mantener a empleados y usuarios alerta frente a las amenazas emergentes de los deepfakes. Las empresas deben organizar sesiones periódicas de capacitación en ciberseguridad, simulacros de fraude y talleres para identificar señales de alerta, como cambios sutiles en el comportamiento de los interlocutores o inconsistencias en la comunicación. Mantenerse actualizado sobre las nuevas técnicas de manipulación y los avances en detección permite reaccionar con rapidez y eficacia ante posibles incidentes, reduciendo así el impacto de los fraudes y mejorando la capacidad de respuesta de toda la organización.

Desafíos futuros y perspectivas

La tecnología de deepfakes está evolucionando a un ritmo sin precedentes, impulsada por la democratización de la inteligencia artificial y la proliferación de herramientas capaces de generar videos, audios e imágenes cada vez más realistas y difíciles de distinguir de los originales. Desde 2022, el desarrollo de modelos avanzados de conversión de texto en video por parte de empresas como OpenAI, Google y otras ha supuesto un salto cualitativo, facilitando la creación masiva de contenidos sintéticos tanto para usos legítimos como maliciosos. Este avance ha provocado un aumento significativo de los casos de fraude y desinformación: solo en el primer trimestre de 2025, los incidentes de deepfakes superaron el total registrado en todo el año anterior, alcanzando aproximadamente el 7% de la actividad fraudulenta global. Además, la inteligencia artificial potencia otras formas de ciberataque, como el phishing, permitiendo a los atacantes personalizar y automatizar sus estrategias a gran escala y con mayor eficacia.

La reducción de las barreras técnicas ha hecho que cualquier persona con acceso a internet pueda generar deepfakes sofisticados desde un simple smartphone, lo que dificulta enormemente el control y la verificación fiable del contenido manipulado. Esta accesibilidad, junto con la rápida difusión en redes sociales y plataformas digitales, amplifica el alcance de los fraudes, la desinformación y el daño reputacional, complicando la protección de la identidad y la privacidad tanto de individuos como de organizaciones.

Frente a este panorama, surgen nuevos retos para la ciberseguridad y la sociedad en general. La creciente sofisticación de los deepfakes exige la adopción de tecnologías de detección más avanzadas, como las que analizan microanomalías imperceptibles a simple vista, y la colaboración internacional para el desarrollo de normas y regulaciones que permitan identificar y etiquetar de forma fiable el contenido generado por IA. Además, la formación continua y la concienciación de los usuarios seguirán siendo piezas clave para reconocer señales de manipulación, como movimientos faciales poco naturales o desincronización entre audio e imagen. La lucha contra los deepfakes requerirá, por tanto, una combinación de innovación tecnológica, cooperación global y una cultura digital más crítica y vigilante.

Referencias

A continuación se adjuntan todos los documentos, artículos y recursos empleados para la realización de este informe:

Noticias:

- Noticia sobre el uso de GANs en la película “The Irishman”: https://www.bbc.com/mundo/noticias-50234735
- Noticia sobre empleado estafado con deepfake: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk
- Noticia sobre la suplantación de Brad Pitt: https://www.infobae.com/espana/2025/01/13/una-mujer-pierde-casi-un-millon-al-pensar-que-tenia-una-relacion-con-brad-pitt-los-estafadores-se-hicieron-pasar-por-el-actor-con-inteligencia-artificial/
- Noticias sobre la suplantación de Elon Musk: https://economictimes.indiatimes.com/tech/technology/how-deepfake-musk-became-internets-biggest-scammer/articleshow/112618750.cms?from=mdr
- Noticia sobre la suplantación de Biden: https://es.wired.com/articulos/audio-falso-de-biden-es-el-principio-de-la-desinformacion-politica-generada-por-ia
- Noticia sobre el intento de fraude a Ferrari: https://www.xataka.com/robotica-e-ia/ceo-ferrari-ha-sido-ultima-victima-deepfakes-para-estafas-a-empresa-pudo-salirle-muy-caro

Estadísticas sobre deepfakes:

- https://www.itdigitalsecurity.es/endpoint/2025/02/los-intentos-de-fraude-con-deepfakes-han-crecido-un-2137-en-los-ultimos-tres-anos
- https://www.escudodigital.com/ciberseguridad/deepfakes-famosos-disparan-2025-trump-mas-suplantado_63063_102.html
- https://www.feedzai.com/es/pressrelease/tendencias-de-fraude-con-ia-2025/
- https://es.statista.com/grafico/31907/paises-con-mayores-aumentos-de-casos-de-fraude-de-deepfake/

- https://www.gartner.com/en/newsroom/press-releases/2024-02-01-gartner-predicts-30-percent-of-enterprises-will-consider-identity-verification-and-authentication-solutions-unreliable-in-isolation-due-to-deepfakes-by-2026