DEEPFAKES: RIESGOS, CASOS REALES Y DESAFÍOS EN LA ERA DE LA IA

isms forum GIA | GRUPO DE INTELIGENCIA ARTIFICIAL

MARZO 2025

Autores

PARTICIPANTES	Antonio Fernandes
	Daniel Fernández
	Diana Molero
	Miguel Ángel Pérez
GESTIÓN DEL PROYECTO	Beatriz García
DISEÑO / MAQUETACIÓN	Marta Barroso
	Lydia García

ÍNDICE

1. Introducción

En la era digital actual, la tecnología avanza a pasos agigantados, ofreciendo innumerables beneficios y oportunidades. Sin embargo, también trae consigo nuevos desafíos y amenazas.

Uno de los fenómenos que han surgido con la Inteligencia Artificial y que plantean un gran reto son los deepfakes. Estas sofisticadas falsificaciones digitales tienen el potencial de engañar a millones de personas al presentar imágenes, videos o audios falsos que parecen increíblemente reales.

Los ciberdelincuentes están utilizando campañas de desinformación y contenido de deepfake para desinformar al público sobre eventos, influir en la política y las elecciones, contribuir al fraude y manipular a los accionistas en un contexto corporativo. Muchas organizaciones han comenzado a ver los deepfakes como un riesgo potencial aún mayor que el robo de identidad. Esta preocupación se reflejaba ya en un informe de University College London (UCL) publicado en 2020, que clasificaba la tecnología de deepfake como una de las mayores amenazas que enfrenta la sociedad hoy en día.

Estas amenazas tienen aún más riesgo, cuando cada vez la implementación de los deepfakes es más sencilla con el uso de aplicaciones de fácil manejo, convirtiéndose en herramientas cuyo uso se ha democratizado, herramientas para "todos los públicos". Pero no sólo eso, sino que ya los deepfakes con objetivos criminales se han convertido en un servicio emergente, Deepfake as a Service, en el que actualmente algunos ciberdelincuentes disponen de un portfolio de ejemplos deepfakes de gente conocida, y muestran el mismo para que su potencial cliente pueda hacerse una idea de lo que el artista puede crear, oscilando los precios entre los 10$ para imágenes y 500$ por minutos de vídeo.

En este documento, el objetivo principal es presentar el Observatorio de Deepfakes para exponer y analizar los casos de deepfakes más relevantes ocurridos hasta 2024; así como sus repercusiones, elaborado desde el Grupo de Inteligencia Artificial, en adelante GIA, de ISMS Forum.

Para ello en primer lugar se facilitará una introducción sobre deepfakes, y los diferentes tipos en lo que se pueden clasificar, seguidos del análisis de los casos de más relevantes, para dar paso finalmente a la propuesta de "mejores prácticas" para facilitar la detección, acompañadas por algunas herramientas emergentes.

A partir del análisis de los casos de deepfakes expuestos y las guías y herramientas para su detección, se espera posicionar al lector en una situación de menor nivel de exposición al riesgo frente a posibles deepfakes malintencionados.

Este enfoque facilita que la tecnología deepfake se utilice de manera ética y responsable, fomentando la innovación y manteniendo la confianza pública en los avances de la IA.

2. ¿Qué es un deepfake?

El término "deepfake" es una combinación de "deep learning" (aprendizaje profundo) y "fake" (falso), haciendo referencia a la **fusión de inteligencia artificial con medios manipulados**, creando **falsificaciones convincentes en imágenes, vídeos o grabaciones de audio**. Utilizando el aprendizaje automático avanzado (ML), los deepfakes emulan meticulosamente a personas reales, ampliando los límites de nuestra capacidad para distinguir entre contenido digital auténtico y manipulado.

En concreto, el término "deepfake" se originó en 2017, para referirse a un tipo específico de medios sintéticos en los que el parecido de una persona se superpone al cuerpo de otra o en diversos escenarios, a menudo mediante el uso de algoritmos de aprendizaje profundo y técnicas de inteligencia artificial.

Ilustración 1. Imágenes creadas por IA en el experiment "Detect Fakes" experiment, de la Universidad Northwestern's Kellogg School of Management

La evolución de la tecnología deepfake tiene sus raíces en la convergencia de la inteligencia artificial (IA), especialmente el aprendizaje profundo, con amplios conjuntos de datos. Sin embargo, la tecnología detrás de los deepfakes tiene raíces que se remontan más atrás, ya que las técnicas subyacentes, como el aprendizaje profundo y las redes neuronales, han estado en desarrollo durante años. Estas redes neuronales generativas antagónicas (GAN por sus siglas en inglés) utilizan algoritmos que aprenden de patrones presentes en las imágenes, vídeos o audios para después manipularlos y recrear una imagen de una cara, persona o cualquier otro tipo de objeto.

A medida que avanza el aprendizaje automático y aumenta la potencia de cálculo, los deepfakes han evolucionado desde el simple intercambio de rostros hasta la compleja imitación de voces, poniendo en entredicho la autenticidad de los contenidos digitales.

Los deepfakes se utilizan con frecuencia en diversos sectores, pero han cobrado especial importancia en los medios de comunicación, el entretenimiento y la ciberseguridad. El auge de la tecnología deepfake ha suscitado una gran preocupación por su posible uso indebido, en particular para difundir información errónea o crear pornografía no consentida.

Uno de los sectores con mayor nivel de riesgo debido al uso deepfakes fraudulentos, por la relevancia de la identidad, es el sector de la banca y las finanzas. La tecnología deepfake podría utilizar grabaciones de voz y vídeo de clientes para crear comunicaciones fraudulentas casi indistinguibles de las interacciones auténticas. Algunos ejemplos concretos de fraudes basados en el uso de deepfakes en este sector son los siguientes:

Fraude de cuentas nuevas:

Creación de identidades sintéticas para abrir nuevas cuentas, lo que da lugar a actividades como la acumulación de deudas y el blanqueo de dinero.

Identidades sintéticas:

Utilización de credenciales robadas o falsas para construir identidades artificiales, obtener préstamos y adquirir tarjetas de crédito o débito.

Fraudes fantasmas:

Explotación de identidades robadas de personas fallecidas recientemente para violar cuentas, drenar fondos y participar en diversas actividades fraudulentas.

Reclamaciones de muertos vivientes:

Falsificaciones de familiares que convencen a las entidades financieras de que una persona fallecida sigue viva, lo que da lugar al cobro de prestaciones existentes.

Fraude del CEO:

Cada vez son más frecuentes escenarios donde se realizan llamadas falsas o videoconferencias donde el impostor, haciéndose pasar por un ejecutivo conocido o una figura de autoridad, solicita información confidencial o autoriza transacciones financieras.

Todas estas casuísticas plantean importantes retos para las medidas de seguridad, desafiando las más tradicionales, por lo que la detección de estas falsificaciones es crucial para proteger la información personal y financiera.

Otro de los ejemplos más comunes de uso mal intencionado es la manipulación no consensuada mediante el uso de deepfakes, para comprometer a individuos con fines de acoso y chantaje.

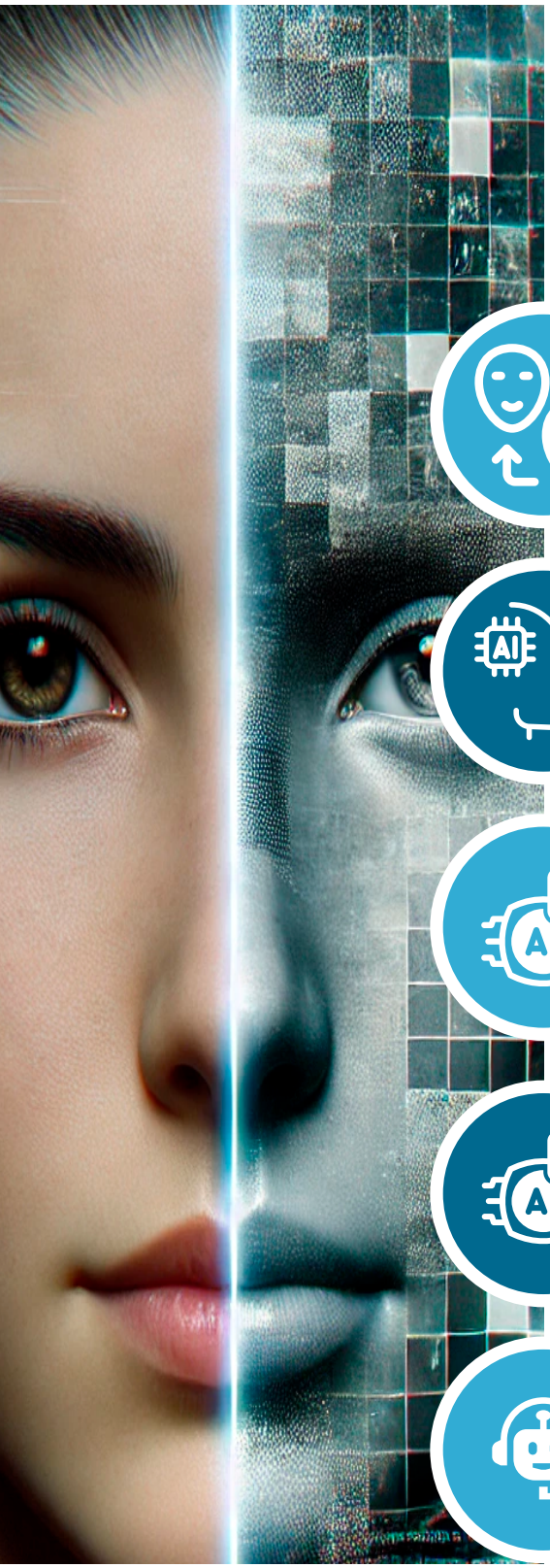
Como último ejemplo relevante, y ante una situación geopolítica tan complicada como la actual, otro de los riesgos es el uso de los deepfakes con fines de desinformación y propaganda. Los deepfakes pueden difundir información falsa, manipular la percepción pública e influir en los resultados políticos al mostrar a figuras políticas haciendo declaraciones falsas o participando en comportamientos poco éticos. Esta táctica puede socavar a los oponentes, influir en la opinión pública y desestabilizar sociedades, lo que lleva a una desinformación generalizada, una erosión de la confianza en las instituciones públicas y disturbios sociales. La naturaleza convincente de los deepfakes representa una amenaza significativa para el discurso político y la estabilidad social.

El rápido incremento de casos de deepfakes para uso fraudulento impulsa la necesidad para cooperar de forma global y combatir estas sofisticadas amenazas.

Esta tecnología emergente, al tiempo que muestra impresionantes avances en IA, plantea problemas éticos y legales. Subraya la necesidad de disponer de técnicas de detección avanzadas y de una normativa que impida su uso indebido para difundir información errónea, manipular los mercados de valores o cometer robos de identidad.

A medida que la tecnología evoluciona, los legisladores y los expertos jurídicos están tratando de encontrar la manera de abordar eficazmente los retos que plantea la tecnología de deepfake, ya que su uso de forma legítima también supone avances en ámbitos como el entretenimiento, creando personajes digitales y reduciendo costes; la industria de la moda, facilitando experiencias personalizadas e incluso la salud, facilitando simulaciones realistas para el ensayo médico.

3.Tipos de deepfakes



Los deepfakes, desde el punto de vista del riesgo que suponen y su complejidad para la detección, pueden ser clasificados por el tipo de contenido manipulado según se detalla a continuación:

Deepfaces

Este tipo de deepfake se centra en la creación de imágenes convincentes, pero completamente falsas creadas desde cero. Utiliza el aprendizaje automático para manipular y generar rostros que parecen reales.

Deepvoices

En este caso, el audio es el elemento manipulado. Se utiliza inteligencia artificial para clonar la voz de una persona y hacer que diga cosas que nunca dijo.

Deepfakes de vídeo

Este tipo suele alterar o reemplazar la cara de una persona en un vídeo, creando imágenes falsas de situaciones que nunca ocurrieron.

Deepfakes de imágenes

Similar a los deepfakes de vídeo, pero aplicados a fotografías o imágenes estáticas.

Deepfakes en tiempo real

Son los deepfakes más complejos donde se utilizan técnicas de vídeo y audio, para la suplantación en tiempo real, haciendo más realista las situaciones y por tanto más difíciles de detectar la manipulación si está bien realizada.

4.Observatorio deepfakes

Observatorio deepfakes

El aumento del número de casos de deepfakes, y sobre todo sus repercusiones e impacto en la sociedad y los retos que supone su detección a nivel de seguridad, han motivado la recopilación en este observatorio y su análisis.

En este capítulo se analizarán los 10 casos de deepfakes seleccionados como parte del observatorio, considerados los más relevantes hasta febrero de 2025, tanto por su tipología, por su impacto económico, reputacional y riesgo elevado de desinformación, como por su rápida evolución y democratización del uso de herramientas de IA, demostrando que los deepfakes pueden ser generados por casi por cualquier persona, como se verá reflejado en algunos de los casos del observatorio.

Para comprender mejor el impacto y los riesgos asociados a estas tecnologías, hemos clasificado los casos recientes de deepfakes según la tecnología utilizada. Esta clasificación nos permite analizar las amenazas específicas de cada tipo de deepfake y entender cómo se propagan en diferentes entornos.

4.1 Deepfake de Video

Estos deepfakes utilizan redes neuronales avanzadas para alterar videos de manera realista, cambiando el rostro, las expresiones y la sincronización labial para hacer que una persona parezca decir o hacer algo que nunca ocurrió.

Caso 1: Deepfake Vladimir Putin

Ficha:

Fecha:	Tecnología usada:	Impacto:
2020	Deepfake de video, con algoritmo OpenSource	Concienciar sobre la desinformación.

Fuentes:

01 02

1. https://www.technologyreview.com/2020/09/29/1009098/ai-deepfake-putin-kim-jong-un-us-election/
2. https://lab.witness.org/projects/synthetic-media-and-deep-fakes/

Descripción:

Video de Vladimir Putin advirtiendo a la población de que puede haber fraude electoral.
La intención era impactar a los espectadores y hacerles reflexionar sobre la importancia de proteger los derechos de voto y combatir la desinformación, concienciando así sobre los peligros de los deepfakes.

Cómo se hizo:

Se hizo sobre un video base en el que hay un actor y se utilizó software libre para intercambiar caras.
El video fue producido utilizando inteligencia artificial para superponer el rostro de Putin.

Sotware:

- Roop: https://github.com/s0md3v/roop
- ReActor: https://github.com/Gourieff/sd-webui-reactor
- https://github.com/iperov/DeepFaceLab

Caso 2: Intento de estafa a WPP

Ficha:

Fecha:	Tecnología usada:	Impacto:
14/05/2024	Deepfake video, clonado de voz	No tuvo impacto por su detección gracias a las medidas de prevención aplicadas, entre ellos la concienciación de las posibles víctimas.

Fuentes:

01 02

Descripción:

WPP, la mayor empresa de publicidad del mundo fue víctima de una estafa de deepfake. Los estafadores crearon una cuenta de WhatsApp utilizando una imagen pública del CEO, Mark Read, y organizaron una reunión por videoconferencia que simulaba involucrar tanto a él como a otro ejecutivo de alto rango.
El intento de fraude fue detectado a tiempo y el CEP de WPP envió un email a sus empleados para concienciar y mantenerlos alerta.

1. Https://www.marketingdirecto.com/digital-general/digital/mark-read-ceo-de-wpp-victima-de-una-estafa-de-deepfake-le-clonaron-incluso-la-voz
2. https://www.theguardian.com/technology/article/2024/may/10/ceo-wpp-deepfake-scam

Puntos clave del caso:

- **Actores:**
  Víctima: Empleados de otra agencia publicidad
  Suplantación de CEO WPP Mark Read y otro ejecutivo
- **Objetivos:** Conseguir dinero y datos personales
- **Medios empleados:**
  Crearon cuenta de WhatsApp con la imagen del CEO
  Utilizaron una clonación de voz y vídeos de YouTube para hacer que la reunión pareciera auténtica.

Medidas preventivas/recomendaciones:

1. **Sospechas durante la videollamada**: Durante la reunión en Microsoft Teams, algunos empleados notaron inconsistencias en la voz y el comportamiento de los supuestos ejecutivos. La voz clonada y los vídeos de YouTube utilizados no lograron replicar perfectamente las características naturales de los ejecutivos.

2. **Verificación de identidad**: Los empleados verificaron la identidad de los ejecutivos a través de otros canales de comunicación. Al darse cuenta de que los verdaderos ejecutivos no estaban al tanto de la reunión, se confirmó que se trataba de un intento de estafa.

3. **Alertas internas**: WPP había implementado protocolos de seguridad que incluían la verificación de solicitudes inusuales, como la creación de nuevas empresas o transferencias de dinero. Al darse cuenta de que los verdaderos ejecutivos no estaban al tanto de la reunión, se confirmó que se trataba de un intento de estafa.

Caso 3: Estafa impersonando a la Princesa Leonor

Ficha:

Fecha:	Tecnología usada:	Impacto:
07/12/2024	Deepfake video y audio grabados	245€ por estafa

Fuentes:

01 02

1. https://www.newtral.es/tiktok-estafa-leonor-te-lo-explicamos/20241209/ 2. https://lab.witness.org/projects/synthetic-media-and-deep-fakes/
2. https://www.tiktok.com/@leonorprincesa__/video/7434256847734983991

Descripción:

Se han compartido vídeos con diferentes modalidades de estafa, pero lo más habitual es que aparezca un deepfake de la princesa Leonor instando a los usuarios a capturar la imagen de un dibujo cuando encaje correctamente. Tras ello, algunos estafadores solicitan a los usuarios comentar la publicación o escribir un mensaje privado diciendo que lo han conseguido para ponerse en contacto con ellos.

Otra modalidad consiste en un vídeo o audio que suplanta a la princesa Leonor y dice a los usuarios que le escriban un mensaje a través de TikTok con el dinero que necesitan y ahí comienza la estafa.

Según informa El País, para recibir el supuesto premio los estafadores solicitan a la víctima un ingreso en concepto de "depósito", tras ello, le siguen solicitando dinero hasta que desaparecen.

El medio citado contactó con Juana Cobo, una afectada de esta estafa de TikTok que suplanta a la princesa Leonor. Cobo relató que recibió un mensaje en la red social en el que supuestamente le hablaba la princesa Leonor diciendo que había ganado 100.000 dólares, pero que tenía que pagar un impuesto de 2.200 quetzales guatemaltecos (unos 245 euros) para conseguir el dinero.

Al realizar el primer pago, relata El País, los estafadores solicitaron más dinero hasta que Cobo se dio cuenta del timo y les dijo que eran unos estafadores. Entonces, la bloquearon y desaparecieron.

Caso 4: Estafa de criptomonedas con deepfake de Elon Musk

Ficha:

Fecha:	Tecnología usada:	Impacto:
De Marzo a Mayo 2024	Deepfake video	Pérdidas financieras, estimándose en USA 80 millones de dólares en criptomonedas, pérdida de confianza en las plataformas como Tiktok y Youtube, reputación de figuras públicas, aumento vigilancia en redes sociales.

Fuentes:

01 02 03

1. https://www.elgrupoinformatico.com/seguridad/estafa-deepfake-elon-musk-tiktok-criptomonedas/
2. https://www.adslzone.net/noticias/seguridad/youtube-deepfakes-elon-musk-regalan-criptomonedas/
3. https://es.cointelegraph.com/news/yikes-elon-musk-warns-users-against-latest-deepfake-crypto-scam

Descripción:

Durante 2024 ha habido varios casos de estafas de criptomonedas utilizando deepfakes de Elon Musk. Los estafadores crean videos falsos muy realistas en los que Musk aparentemente regala criptomonedas, pero en realidad buscan engañar a las personas para que entreguen sus propias criptomonedas.desaparecieron.

Por ejemplo, en YouTube se han publicado videos en los que Musk supuestamente regala criptomonedas a cambio de que los usuarios transfieran las suyas a una dirección específica. Estos videos pueden parecer muy convincentes, especialmente cuando se transmiten en vivo y utilizan cuentas verificadas.

En TikTok también se han visto estafas similares, donde los deepfakes de Musk promocionan sorteos de criptomonedas falsos. Los estafadores piden a las víctimas que se registren en plataformas fraudulentas como Moonexio, Algetxio, Cratopex o BitVex, y depositen una pequeña cantidad de criptomonedas para recibir una mayor cantidad a cambio, lo cual nunca ocurre.

¿Cómo se hizo?

Los estafadores han utilizado principalmente dos tipos de deepfakes en las estafas de criptomonedas con la imagen de Elon Musk:

- Simulaciones de videos de Elon Musk en tiempo real mediante transmisiones en vivo en cuentas verificadas.

- Entrevistas manipuladas, clonando la voz y los gestos, tomando muestras de imágenes originales de una charla TED en la participó Musk, junto con entrevistas y otras exposiciones mediáticas.

Elon Musk ha tomado varias medias para combatir las estafas que utilizan deepfakes de su imagen:

- Advertencias públicas a través de sus redes sociales, como Twitter desmintiendo públicamente los vídeos y alertando a los usuarios.

- Colaboración con plataformas afectadas como TouTube y TikTok para eliminar los videos deepfake.

- Promoción de la seguridad en redes sociales: Desde que adquirió Twitter, Musk ha enfatizado la importancia de eliminar bots de spam y cuentas fraudulentas. Ha prometido mejorar la seguridad en la plataforma para proteger a los usuarios de estafas y otros tipos de fraude.

4.2 Deepfake de Audio

El uso de modelos de IA capaces de clonar la voz de una persona permite la creación de audios realistas que pueden utilizarse para fraudes, manipular declaraciones o engañar a familiares y empleados.

Caso 5: Deepfake racista en un colegio en Estados Unidos

Ficha:

Fecha:	Tecnología usada:	Impacto:
01/2024	Deepfake audio	Desinformación y daño reputacional, despido del profesor suplantado

Fuentes: 01 02 03

Descripción:

El audio deepfake del director de una escuela provoca amenazas de muerte en Maryland, EE. UU. El clip de audio falso, creado con inteligencia artificial, mostraba a un director de escuela haciendo comentarios racistas y antisemitas. Este clip se volvió viral, provocando amenazas de muerte contra el director y causando una gran división en una comunidad cerca de Baltimore, con grandes poblaciones negras y judías. Sin embargo, se descubrió que el clip era falso y había sido manipulado por IA.

La policía identificó y arrestó a Dazhon Darien, director de atletismo de la escuela, como el responsable de crear el clip falso, atribuyéndole cargos de represalias contra un testigo y acoso. Se alega que Darien, bajo investigación por un presunto robo, creó el clip para desacreditar al director antes de ser despedido.

Para crear el deepfake, Darien utilizó herramientas de inteligencia artificial que permiten la clonación de voz. Estas herramientas funcionan identificando patrones de voz únicos de una persona y luego reproduciéndolos para leer cualquier texto, pudiendo hacerse con solo unos segundos de grabación de la voz original. Destacar el fácil acceso a las herramientas de deepfakes de voz, pues utilizó la red informática de las Escuelas Públicas del Condado de Baltimore para acceder a herramientas de IA.

1. https://www.bbc.com/news/articles/ckg9k5dv1zdo
2. https://incode.com/blog/top-5-cases-of-ai-deepfake-fraud-from-2024-exposed/
3. https://theconversation.com/el-peligro-real-de-los-deepfakes-de-clonacion-de-voz-y-como-detectarlos-215012

Debido a la situación de la comunidad donde ocurrió, susceptible a este tipo de incidentes racistas, el audio consiguió mayor credibilidad. Además, se mencionaba jerga y otros detalles como nombres del personal, que solo las personas cercanas a la escuela conocerían.

Sin embargo, cuando se escucha con atención, hay ediciones claras entre las oraciones, y la voz, aunque similar a la del director, suena bastante monótona.

El director de la escuela finalmente cambió de trabajo y trabaja en otra escuela, con los consecuentes daños reputacionales.

El incidente destaca los peligros de la desinformación generada por IA y cómo puede afectar a las comunidades locales. A pesar de las políticas de las redes sociales para limitar la difusión de publicaciones generadas por IA, estas acciones solo se toman cuando se puede probar que un clip es falso, momento en el cual ya podría haber alcanzado a millones de personas.

Caso 6: Llamadas a los votantes suplantando a Biden en elecciones USA 2024

Ficha:

Fecha:	Tecnología usada:	Impacto:
22/01/2024	Deepfake audio	Desinformación, fraude electoral Condena de 6M$ para el autor del deepfake y 2$ para la operadora que lo difundió

Fuentes: 01 02 03 04

Descripción:

El caso de una llamada automática generada por IA que imita a Joe Biden y anima a los demócratas a no votar en las primarias demócratas de New Hampshire en enero es solo uno de los innumerables ejemplos de deepfakes utilizados para cometer fraude electoral.

En respuesta, ha habido un impulso continuo para que los gobiernos regulen el uso de dicha tecnología en las campañas políticas, dado lo susceptibles que son los votantes a la desinformación. En los EE. UU., el grupo de defensa Public Citizen pidió a la Comisión Federal de Elecciones (FEC) que regule el uso de la IA en los anuncios de campaña.

1 https://www.reuters.com/world/us/fake-biden-robo-call-tells-new-hampshire-voters-stay-home-2024-01-22/
2. https://www.fcc.gov/document/fcc-eb-settles-lingo-transmitting-illegal-robocalls
3. https://www.nbcnews.com/politics/politics-news/steve-kramer-admitted-deepfaking-bidens-voice-new-hampshire-primary-rcna153626
4. https://docs.fcc.gov/public/attachments/DOC-405811A1.pdf

El deepfake de New Hampshire es un recordatorio de las muchas formas en que los deepfakes pueden sembrar confusión y perpetuar el fraude.

En agosto de 2024, Lingo Telecom, el proveedor de servicios de voz que distribuyó las llamadas automáticas generadas por inteligencia artificial a través de números de teléfono "falsificados", acordó pagar una multa de 1 millón de dólares por su papel en la estafa del deepfake de Joe Biden. La combinación potencial del uso indebido de la tecnología de clonación de voz generativa de IA, y la suplantación de identidad de llamadas (Caller ID spoofing) en la red de comunicaciones de EE. UU. presenta una amenaza significativa. La Comisión Federal de Elecciones considera que los proveedores de servicios de comunicaciones son la primera línea de defensa contra estas amenazas y serán responsables de garantizar que hagan su parte para proteger al público estadounidense.

Por otro lado, Steve Kramer, el consultor político que trabajaba para la campaña de un demócrata, admitió que estaba detrás del fraude a través de deepfake de audio y fue imputado con una multa de 6M$ y 26 delitos graves de supresión del voto y de suplantación de un candidato. Kramer encargó el deepfake al mago Paul Carpenter, quien utilizó un software muy económico, indicando que el proceso sólo costó un dólar y realizándolo en menos de 20 minutos de trabajo.

4.3 Imágenes Generadas con IA

Las herramientas de generación de imágenes permiten crear rostros falsos extremadamente realistas, lo que facilita la difusión de contenido engañoso y la manipulación de la opinión pública.

Caso 7: Imágenes sexualizadas de Taylor Swift

Ficha:

Fecha:	Tecnología usada:	Impacto:
25/1/2024	Deepfake de imagen con software comercial trucado	Daño moral

Fuentes: 01 02

1.https://edition.cnn.com/2024/01/25/tech/taylor-swift-ai-generated-images/index.html
2. https://indianexpress.com/article/technology/artificial-intelligence/taylor-swift-ai-deep-fake-explicit-images-9130006/

Descripción:

Se detectaron imágenes de Taylor Swift desnuda y sexualizada en Twitter, las imágenes fueron compartidas millones de veces en la plataforma.

Uno de los posts más compartido tenía más de 45 millones de visualizaciones y más de 24 mil retweets.

Origen:

Las imágenes salieron de un grupo de Telegram donde había gente dedicada a generar imágenes de desnudos de mujeres, una de las herramientas usadas fue el generador de imágenes gratuito de texto a imagen de Microsoft.

Es importante destacar que estas imágenes no son realmente "deepfakes" si nos basamos en la definición original de la palabra. Originalmente, un deepfake se refería a imágenes o videos creados utilizando redes adversariales entrenadas en un rostro para reemplazarlo en otro cuerpo. Básicamente, en lugar de usar IA para superponer el rostro de Taylor Swift en una imagen pornográfica real, estas fueron creadas desde cero usando IA generativa.

¿Cómo se hizo?

De una forma similar a como se hackean LLMs, utilizando palabras clave para "despistar" al sistema de filtrado que evita usar imágenes de famosos.

Metiendo palabras extras en los nombres de las celebridades o alterando el orden o jugando con la sintaxis.

Caso 8: IA para crear material de abuso sexual infantil

Ficha:

Fecha:	Tecnología usada:	Impacto:
28/02/2025	Deepfake video y audio	Daños sociales con la normalización de la explotación infantil. Pendiente de imputación de los detenidos y las condenas correspondientes.

Fuentes: 01 02

Descripción:

Europol anunció la desarticulación de una red internacional dedicada a la creación y distribución de imágenes de pornografía infantil generada por inteligencia artificial (IA). Se han detenido a 25 miembros de la organización criminal, dos de ellos ubicados en España. La investigación comenzó a partir de la detención en el año 2024 de un ciudadano danés, quien dirigía una plataforma online a través de la que se distribuía material generado por IA que producía previamente. Usuarios de todo el mundo, después de realizar un pago simbólico en línea, podían obtener una contraseña para acceder al material de esta plataforma donde podían visualizar abusos a menores.

Los delincuentes se sirven de modelos de IA capaces de generar o alterar imágenes para producir pornografía infantil y extorsión sexual. Estos modelos están ampliamente disponibles y se han desarrollado de forma muy rápida con resultados que actualmente se parecen cada vez más a las imágenes reales. En muchos casos cuentan con bases de datos de miles de imágenes delictivas y/o de menores en fuentes en línea, o vídeos superponiendo dichas imágenes, con los que entrenan los modelos para generar las nuevas imágenes.

Esto plantea importantes desafíos a las autoridades a la hora de identificar a las verdaderas víctimas y a los agresores. Incluso en los casos en los que el contenido es íntegramente generado por IA y no se representa a ninguna víctima real, como en esta operación, sigue contribuyendo a la cosificación y sexualización de los niños.

1. https://www.newtral.es/tiktok-estafa-leonor-te-lo-explicamos/20241209/
2. https://www.tiktok.com/@leonorprincesa_/video/7434256847734983991

4.4 FaceSwap en Tiempo Real

Esta tecnología permite reemplazar el rostro de una persona en tiempo real, lo que facilita fraudes en videollamadas y engaños personalizados.

Caso 9: Estafa del amor usando intercambio de caras desde teléfono

Ficha:

Fecha:	Tecnología usada:	Impacto:
2022	Multiple software	Robo de $652.544.805 de víctimas en 2023 (sobre $82M más que en 2022)

Fuentes: 01

Descripción:

Una estafa del amor es un tipo de fraude en que te haces pasar por otra persona para falsear una relación con la víctima y así engañarla para obtener dinero y datos personales.

Los primeros casos de este tipo de estafas datan del 2022 y utilizan DeepFakes y FaceSwap para mostrar una cara falsa en tiempo real a una persona con la que conversas.

¿Cómo se hizo?

Lo hacen a través de redes sociales y servicios de mensajería desde móviles y portátiles utilizando múltiples programas para de face-swapping tanto en llamadas en tiempo real como posts que dejan en TikTok, Instagram y Facebook entre otras.

1 https://www.gmal.co.uk/realtime-deepfake-dating-scams/

Caso 10: Estafa empleado Arup, empresa de ingeniería británica

Ficha:

Fecha:	Tecnología usada:	Impacto:
16/5/2024	Video en tiempo real	Robo de $25M

Fuentes:

01

Descripción:

Arup, una empresa multinacional británica de diseño e ingeniería detrás de edificios mundialmente famosos como la Ópera de Sydney ha confirmado que fue el objetivo de una estafa que llevó a uno de sus empleados de Hong Kong a pagar 25 millones de dólares a estafadores.

La policía de Hong Kong reveló en febrero que, durante una estafa elaborada, un empleado del sector financiero fue engañado para asistir a una videollamada con personas que creía eran el director financiero y otros compañeros de trabajo. Sin embargo, todos resultaron ser recreaciones "deepfake". En ese momento, las autoridades no revelaron el nombre de la empresa ni de las partes involucradas.

Según la policía, el empleado había sospechado inicialmente de un correo electrónico que parecía un intento de phishing proveniente de la oficina de la empresa en el Reino Unido, ya que solicitaba realizar una transacción secreta. No obstante, el trabajador dejó de lado sus dudas después de la videollamada, ya que las personas presentes se veían y sonaban como colegas que él reconocía.

El empleado, confiado, aceptó transferir un total de 200 millones de dólares de Hong Kong (aproximadamente 25.6 millones de dólares estadounidenses). El monto fue enviado en 15 transacciones, informó la emisora pública RTHK, citando a la policía.

1. https://edition.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk/

5. Análisis de los casos observados

Tras la descripción del Top 10 de casos de deepfakes más representativos ocurridos hasta 2025, a continuación, se reflejarán las principales conclusiones obtenidas del análisis de los mismos, identificando cinco factores clave.

1. Tipología de deepfakes

Los deepfakes analizados, según se describía previamente, por el tipo de tecnologías de IA empleadas y su presentación, pueden clasificarse de la siguiente manera:

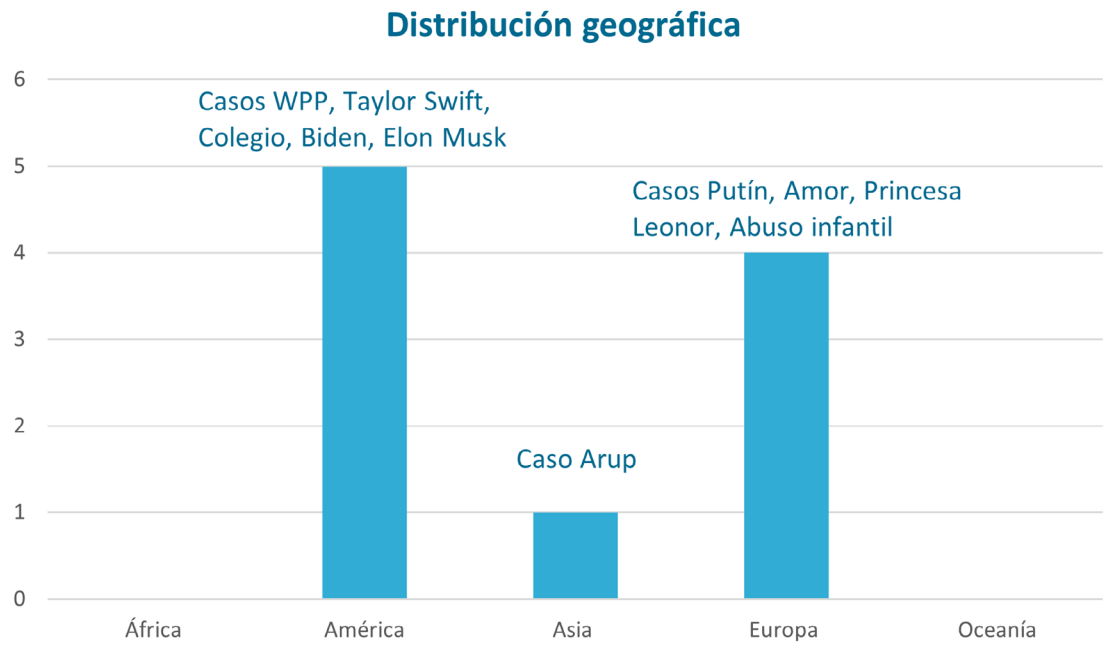
Caso deepfake/ Tipología	Imagen	Audio	Video	Video en tiempo real
Caso 1: Deepfake Vladimir Putin	■		■	
Caso 2: Intento de estafa a WPP				■
Caso 3: Estafa impersonando a la Princesa			■	
Caso 4: Estafa criptomonedas Elon Musk			■	
Caso 5: Deepfake racista en colegio USA		■		■
Caso 6: Fraude elecciones USA deepfake Biden		■		
Caso 7: Imágenes Taylor Swift	■			
Caso 8: IA para crear material abuso infantil			■	
Caso 9: Estafa del amor	■		■	■
Caso10: Estafa empleado Arup				■

Inicialmente simplemente se hacía uso de imágenes reemplazando la cara y/o el cuerpo de la persona suplantada; sin embargo, la evolución y perfeccionamiento de las herramientas de IA en términos de voz y de video ya han conseguido hacer totalmente realistas los vídeos en tiempo real con uso en videollamadas como el caso: estafa empleada Arup. Las técnicas de detección más sencillas, como movimiento de la cabeza, comisura de los labios, parpadeo de los ojos, ya no son válidas ante estos escenarios de estafa, especialmente en ataques dirigidos tan trabajados, y son necesarias herramientas específicas de detección de deepfakes.

Por otro lado, también destacar como en el caso de la desinformación, y con un objetivo de llegar a las masas, todavía los deepfakes de audio resultarán efectivos como en los casos de fraude en las elecciones de USA, con el deep fake de Biden, y el caso del deep fake racista en el Colegoio de USA, pues el usuario sólo tiene que desconfiar de un factor cómo la voz para que resulte creíble, en lugar de voz e imágenes acompasadas con gestos y movimientos.

2. Distribución geográfica

El análisis de la distribución geográfica de los casos documentados revela una concentración significativa en determinadas regiones del mundo. Se observa un alto número de incidentes en países con elevada penetración digital y acceso a tecnologías avanzadas de inteligencia artificial. En particular, Estados Unidos, Europa y China destacan como zonas con mayor actividad reportada, tanto en el uso de deepfakes con fines maliciosos como en la implementación de herramientas de detección.



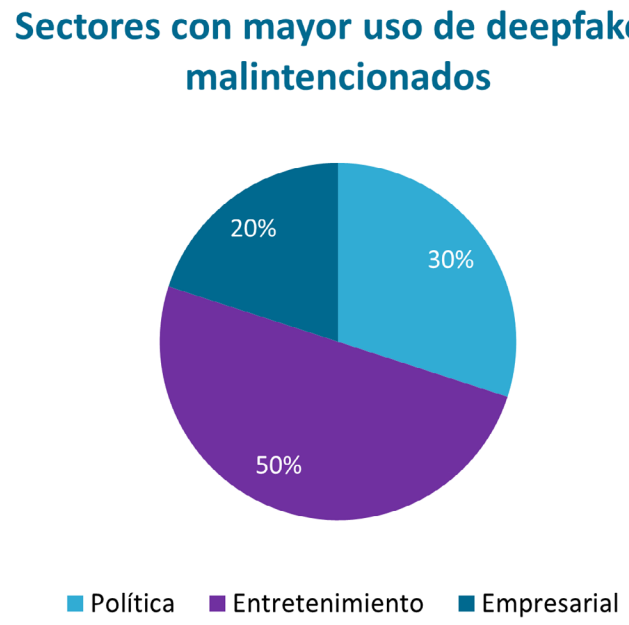
3. Sectores con mayor uso de deepfakes malintencionados

En este caso se analizan en qué ámbitos o sectores el uso de deepfakes suponen una mayor amenazada. La gráfica muestra la distribución porcentual del uso de deepfakes con fines malintencionados en tres sectores clave: política, entretenimiento y empresarial.

El uso de deepfakes con fines malintencionados es una amenaza creciente, con un impacto particularmente fuerte en el entretenimiento y la política, considerando dentro del entretenimiento las redes sociales, precisamente por la capacidad de las mismas para viralizar el contenido.

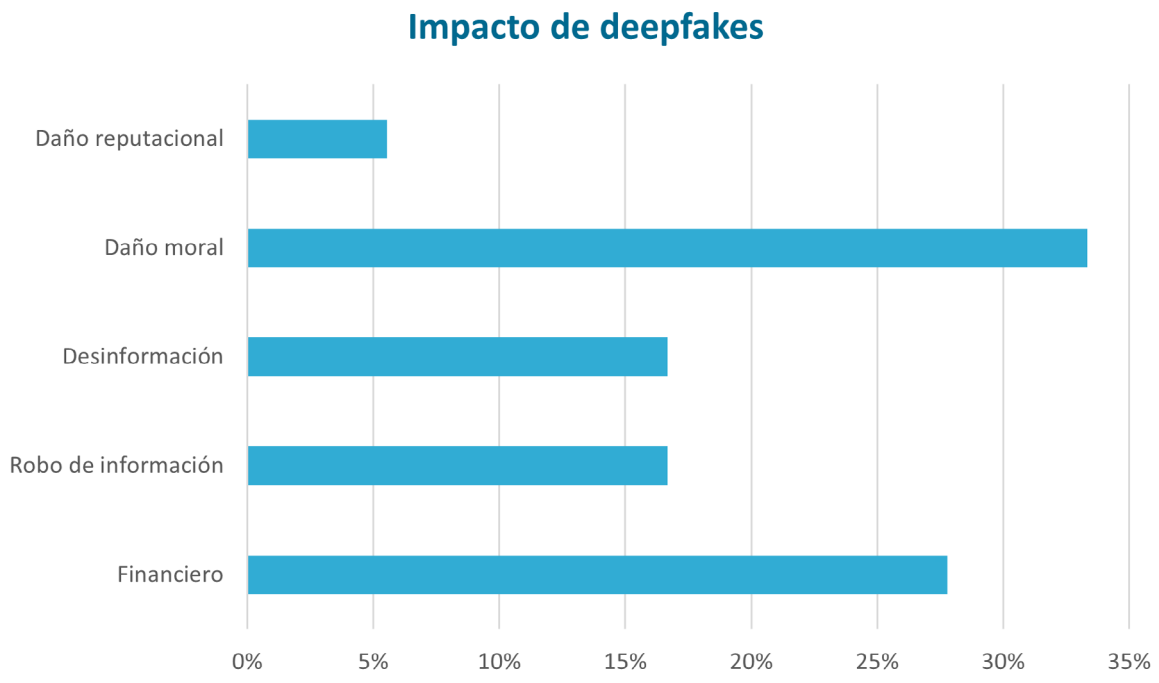
Tanto en el entretenimiento como en la política, las principales amenazas son la desinformación o daño reputacional, pero crecen los ataques masivos con objetivos de fraude financiero en redes sociales, como los ejemplos de la estafa del amor, o la estafa de suplantación de Elon Musk para conseguir criptomonedas.

La facilidad de acceso a herramientas avanzadas de inteligencia artificial ha facilitado la generación de contenido falso, lo que subraya la necesidad de estrategias de detección y regulación para mitigar sus efectos.



4. Impacto

El impacto de los deepfakes se extiende a diversas áreas, siendo el daño moral y el perjuicio financiero los más significativos. Según la gráfica, el daño moral representa el mayor porcentaje, evidenciando el impacto psicológico y emocional que estas manipulaciones pueden causar en las víctimas. El fraude financiero también ocupa una proporción considerable, demostrando el uso creciente de esta tecnología para estafas y suplantaciones de identidad. La desinformación y el robo de información siguen siendo amenazas preocupantes, facilitando la propagación de noticias falsas y la vulnerabilidad de datos sensibles. Aunque el daño reputacional aparece con menor incidencia, sigue siendo un factor relevante, especialmente para figuras públicas y empresas.

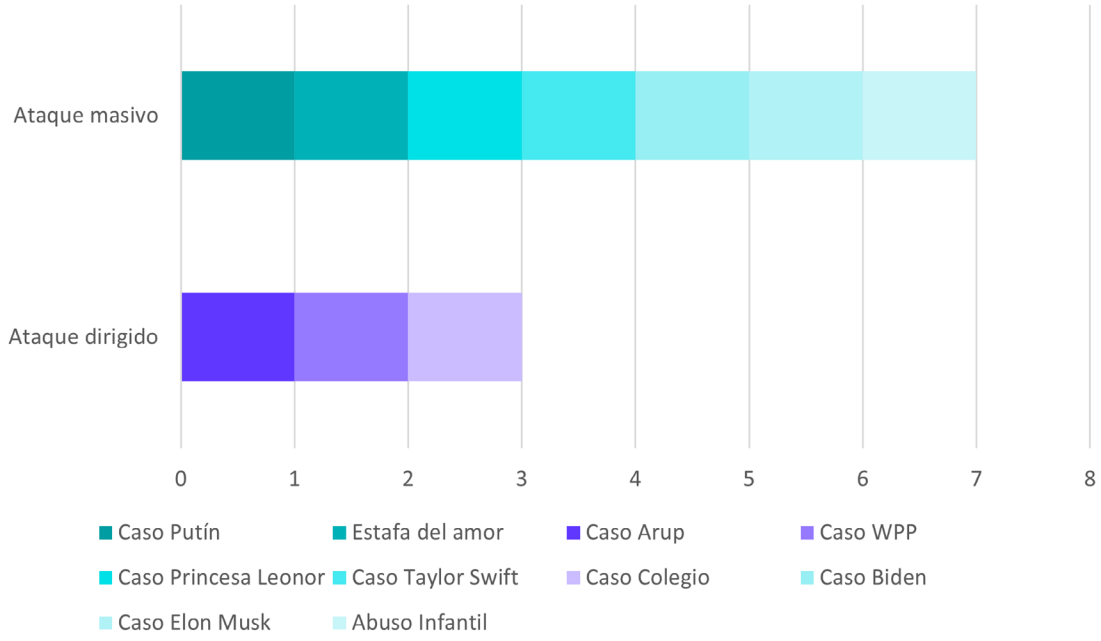


5. Objetivos de ataque

La relevancia de la proporción de ataques masivos frente a ataques dirigidos de un 70% frente a un 30%, pone de manifiesto, una vez más, el riesgo que suponen los deepfakes malintencionados para la población en general.

La concienciación y educación pública serán imprescindibles desde dos ámbitos para gestionar estas amenazas. Por un lado, para que los ciudadanos puedan desconfiar y detectar deepfakes de baja calidad y así reducir el impacto que estos ataques de desinformación, estafa o daño moral y reputacional puedan tener como objetivo. Por otro lado, para hacer uso de tecnología de IA, en concreto de deepfakes, de forma ética y responsable y el fácil acceso a estas herramientas se centre en desarrollar avances en marketing, entretenimiento, educación, atención médica, etc., mejorando las experiencias de los usuarios e impulsando la innovación.

En el caso de los ataques dirigidos, todo parece indicar un mayor incremento de los mismos en los próximos años de la mano de las mejoras en los modelos de entrenamiento y especialmente en el perfeccionamiento de las herramientas de generación de vídeos en tiempo real basadas en IA. En este caso, además de la concienciación y regulación, los avances en métodos y herramientas de detección de deepfakes en los que se suplante la identidad de personas concretamente seleccionadas en tiempo real, se harán imprescindibles para poder parar este tipo de amenazas y lograr así que las empresas especialmente, no se vean comprometidas y sean estafadas.



Este análisis pone de manifiesto la creciente sofisticación y proliferación de esta tecnología en diversos sectores. A medida que su uso se expande, también lo hacen los desafíos asociados a su detección y mitigación. La evolución del panorama de los deepfakes exige una mayor concienciación, la aplicación de técnicas avanzadas de análisis y la cooperación entre sectores públicos y privados para reducir su impacto negativo en la sociedad.

6. ¿Cómo detectar deepfakes?

La detección de deepfakes supone desafíos importantes, también en el ámbito tecnológico. Algunos aspectos clave son:

- **Dependencia del conjunto de datos**: Los algoritmos de detección se entrenan en conjuntos de datos específicos. Una ligera alteración del método utilizado para generar el deepfake puede, por lo tanto, evitar la detección.
- **Adaptabilidad de los GAN**: Las Redes Generativas Adversarias (GAN) pueden actualizar sus modelos discriminativos para eludir los sistemas de detección.
- **Problemas de compresión**: La compresión de videos o la reducción de tamaño pueden difucultar las inconsistencias en las que se basan los sistemas de detección.
- **Manipulación de bases de datos**: Las bases de datos pueden manipularse para clasificar erróneamente imágenes añadiendo identificadores específicos a una pequeña parte del conjunto de datos.
- **Mejora de la calidad**: El aumento de las capacidades de forense de imágenes y detección de deepfakes impulsa la mayor calidad de los videos deepfake. A medida que mejoran las capacidades de detección, también lo hace la calidad de los deepfakes, con los GAN aprendiendo continuamente a producir falsificaciones más convincentes.

En este capítulo se detallarán algunas de las principales técnicas de detección de deepfakes analizadas y probadas; asía como herramientas basadas en software libre para apoyar esa detección.

6.1 Técnicas de detección de deepfakes

Para la detección de imágenes falsas se utilizan diversas técnicas entra las que podemos destacar:

6.1.1 ELA (Error Level Analysis):

¿Qué hace ELA?

Imagina que una foto es como un dibujo pintado en una hoja de papel. Si alguien borra una parte y la redibuja, esa zona quedará ligeramente diferente al resto (quizás con otro tipo de lápiz o textura). ELA es como una “lupa” que resalta esas diferencias, mostrando qué partes de la imagen podrían haber sido modificadas.

¿Por qué funciona?

- **Compresión de fotos:**

Cuando guardas una foto (especialmente en formato JPEG), la cámara o el programa "comprime" la imagen para que ocupe menos espacio. Es como meter ropa en una maleta: cuanto más aprietas, más cosas caben, pero algunos detalles se arrugan.

- **Cada vez que guardas la foto**, se "arrugan" (pierden detalles) de manera uniforme.
- **Si alguien edita la foto:**

Al modificar una zona (por ejemplo, borrar un objeto), esa parte se guarda de nuevo con una compresión diferente. Es como si cortaras un parche de otra maleta y lo cosieras en la tuya: el tejido no coincidirá. **ELA detecta esas "costuras"** comparando cómo se "arruga" la foto original al comprimirla otra vez.

¿Cómo se ve el resultado?

1. **Mapa de calor:**

ELA genera una imagen en blanco y negro (o colores) donde las zonas más brillantes indican **mayor nivel de error.**

Ejemplo: Si alguien clona un árbol en un paisaje, ELA mostrará ese árbol más brillante, porque su "arrugado" al comprimir no coincide con el resto.

2. **Limitaciones:**

No es infalible: Algunas ediciones profesionales pueden engañar a ELA, y a veces zonas naturales (como la piel) pueden parecer sospechosas sin serlo.

Resumen:

ELA es una herramienta que busca inconsistencias en cómo una foto se "arruga" al comprimirse, señalando zonas que podrían estar alteradas. No prueba que haya manipulación, pero ayuda a identificar dónde mirar con más atención.

6.1.2 Análisis de metadatos

¿Qué son los metadatos?

Imagina que una foto es como una carta que envías por correo. Además del mensaje escrito (la imagen en sí), la carta tiene datos adjuntos: la fecha, el lugar de envío, el sello postal, etc. Los metadatos son esos "datos adjuntos" que guarda automáticamente toda foto digital.

Ejemplos:

1. Fecha y hora exactas en que se tomó.
2. Modelo de cámara o teléfono que la capturó.
3. Ubicación geográfica (si el GPS estaba activo).
4. Si se editó con programas como Photoshop.

¿Cómo ayudan a detectar imágenes falsas?

a) **Inconsistencias en la "historia" de la foto**
Si alguien dice que una foto fue tomada en 2020 con un iPhone 12, pero los metadatos muestran que se creó en 2023 con una cámara profesional, es como encontrar un sello postal moderno en una carta supuestamente antigua. Algo no cuadra.

b) **Huellas de edición**
Si la foto fue retocada con Photoshop o cualquier programa de edición, los metadatos suelen guardar un "registro" de esto. Es como si un pintor dejara una firma en un cuadro que intenta hacerse pasar por antiguo.

c) **Ubicaciones imposibles**
Si una foto muestra un paisaje nevado pero los metadatos indican que se tomó en un país tropical en pleno verano, es como encontrar un mapa del Polo Norte en un libro sobre el desierto

c) **Falta de metadatos**
Si los metadatos han sido borrados deliberadamente, es sospechoso. Sería como recibir una carta sin remitente, fecha ni sello: ¿por qué alguien querría ocultar esa información?

Limitaciones

- Se pueden manipular: Así como alguien puede falsificar un pasaporte, también puede editar o borrar metadatos con programas especiales.

- No siempre están presentes: Al subir fotos a redes sociales (como Instagram o WhatsApp), muchas plataformas eliminan los metadatos automáticamente.

Ejemplos prácticos:

Imagina que alguien publica una foto de un platillo exótico diciendo: "Lo cociné hoy en casa". Pero los metadatos revelan que:

- La foto se tomó hace 3 años.
- Fue editada con una app de filtros.
- La ubicación es un restaurante famoso.

Conclusión: La imagen probablemente no es auténtica o fue robada.

Resumen:

El análisis de metadatos es como investigar el "historial secreto" de una foto: quién, cuándo, dónde y cómo se creó o modificó. Si los datos no coinciden con lo que se afirma, hay motivos para sospechar.

Si lo comparamos con la técnica ELA podríamos decir que ELA busca alteraciones visuales, mientras que los metadatos investigan la "biografía" de la imagen.

6.1.3 Análisis de biología humana

¿Qué es el análisis de biología humana?

Los seres humanos tenemos una "firma biológica" única en nuestros gestos, movimientos y características físicas. Los videos o imágenes falsas (como los deepfakes) suelen fallar al imitar estos detalles porque la inteligencia artificial no entiende completamente cómo funciona el cuerpo humano.

El Misterio de los ojos

- **Parpadeo natural:** Los humanos parpadeamos sin pensar, aproximadamente 15-20 veces por minuto. En muchos videos falsos, los ojos no parpadean, lo hacen demasiado o de forma robótica (como un muñeco de cuerda).
Ejemplo: Si una persona en un video habla por minutos sin parpadear, es sospechoso.

- **Reflejos en la pupila:** Los ojos reflejan la luz del entorno (como una ventana o una lámpara). En imágenes falsas, estos reflejos pueden estar mal ubicados, repetidos o ausentes.
Analogía: Es como ver un retrato pintado donde los ojos no coinciden con la escena detrás.

- **Sonrisas falsas:** Cuando sonreímos de verdad, no solo se mueven los labios, sino también los músculos alrededor de los ojos (las "patas de gallo"). En videos falsos, la sonrisa puede verse "pegada" o demasiado perfecta.
Ejemplo: Como cuando un actor malo finge estar feliz, pero su cara no lo transmite.

- **Movimientos incoherentes:** Al hablar, cejas, labios y mejillas se mueven en sincronía. En deepfakes, estos movimientos pueden parecer "desconectados".
Analogía: Es como ver una marioneta donde los hilos se mueven de forma descoordinada

- **Texturas irreales:** La piel humana tiene poros, vellosidades, pecas o arrugas. Las imágenes falsas suelen mostrar piel demasiado lisa, como plástico, o con patrones repetitivos (como un papel tapiz).
Ejemplo: Como comparar una foto real de una naranja (con textura) con un dibujo de una naranja lisa.

- **Luz y sombras:** Si la luz cae sobre una persona desde la izquierda, todas las sombras de su cara deben seguir esa dirección. En videos falsos, las sombras pueden estar mal alineadas o faltar.
Ejemplo: Como si alguien pegara una foto de sí mismo recortada sobre otro fondo, pero la sombra apunta al lado contrario.

El cuerpo en movimiento

- **Respiración:** Al hablar o moverse, el pecho y los hombros suben y bajan ligeramente. En videos falsos, este movimiento puede ser exagerado, rígido o ausente.

- **Gestos con las manos:** Los movimientos naturales de las manos son fluidos y varían en velocidad. En deepfakes, pueden verse robóticos o repetitivos.
Analogía: Es como comparar a un bailarín profesional con un robot programado para mover brazos.

El "Valle inquietante"

A veces, aunque un video falso sea muy bueno, nuestro cerebro detecta algo raro sin saber qué es. Esto se llama "Valle Inquietante": cuando algo parece casi humano, pero no del todo, y nos genera desconfianza.
Ejemplo: Como esos robots que parecen personas, pero te causan escalofríos al mirarlos.

¿Por qué los deepfakes fallan aquí?

La biología humana es increíblemente compleja: cada gesto, músculo o reflejo depende de años de evolución y de un sistema nervioso que la IA aún no replica a la perfección. Los creadores de deepfakes se enfocan en lo grande (como la forma de la cara), pero **los detalles pequeños los delatan.**

Resumen:

Analizar la biología humana para detectar falsificaciones es como **buscar errores** en un disfraz: por muy bueno que sea, siempre habrá una costura mal hecha, un parpadeo fuera de lugar o una sombra que no encaja

6.1.4 Análisis de luces y sombras

La luz en el mundo real: Reglas básicas

En la vida real, la luz se comporta de manera predecible:

Si hay **una fuente de luz** (como el sol, una lámpara o una ventana), todos los objetos en la escena proyectan **sombras y reflejos coherentes** con esa dirección.
Ejemplo: Si te paras frente a una ventana de día, la luz entra por detrás de ti, y tu sombra se proyecta hacia el frente.

Fallos en las imágenes falsas

Cuando alguien edita una foto o crea un deepfake, a menudo **no replica fielmente las leyes de la luz**. Es como dibujar un sol en la esquina de un cuadro, pero olvidar pintar las sombras de los árboles.

Señales clave de inconsistencia:

1. **Dirección de las sombras:**
   a. Si un objeto añadido (como un árbol o una persona) proyecta una sombra hacia la izquierda, pero el resto de la escena tiene sombras hacia la derecha, algo está mal.
   b. *Ejemplo visual imaginario*: Una foto de una persona en la playa al atardecer (el sol está detrás de ella), pero su sombra se proyecta **hacia el sol** en lugar de alejarse.

2. **Intensidad de la luz:**
   a. Si un objeto editado está muy iluminado, pero el entorno está en penumbra, es sospechoso.
   b. *Ejemplo:* Una persona añadida a una foto nocturna parece estar bajo un foco brillante, mientras el resto está oscuro.

3. **Reflejos imposibles:**
   a. Los objetos brillantes (como ojos, botellas o espejos) reflejan el entorno. Si el reflejo no coincide con la escena, es una bandera roja.
   b. *Ejemplo:* En un retrato falso, los ojos de una persona reflejan una habitación con cortinas rojas, pero el fondo real de la foto es un bosque.

4. **Sombras faltantes:**
   a. Si un objeto flota sobre el suelo sin proyectar sombra, o si la sombra es demasiado difusa para la intensidad de la luz, es artificial.
   b. *Ejemplo:* Un avión editado en el cielo no proyecta sombra sobre las nubes debajo, a pesar de un sol brillante.

¿Cómo se realiza la detección?

Los expertos usan herramientas digitales para resaltar las zonas de luz y sombra, pero la idea básica es:

1. Identificar la fuente de luz principal en la escena (¿de dónde viene la luz?).
2. Verificar que todas las sombras y reflejos sigan esa dirección e intensidad.
3. Buscar "rotos" (zonas donde la luz o sombra no encajan).

Analogía:
Es como armar un rompecabezas donde todas las piezas deben encajar. Si una pieza tiene un color o forma que no coincide con las demás, sabes que no pertenece ahí.

- Ejemplo práctico: El selfie de la montaña

1. **Escenario**: Alguien publica una foto de sí mismo en la cima de una montaña nevada, pero sospechas que es falsa.

2. A**nálisis de luces y sombras**:
   - **Fuente de luz**: El sol está a la izquierda en el cielo (las sombras de las rocas apuntan a la derecha).
   - **Inconsistencia**: La sombra de la persona apunta hacia la izquierda, como si el sol estuviera detrás de ella.
   - **Conclusión**: ¡La persona fue editada en la foto! Su sombra contradice la dirección de la luz del entorno real.

Limitaciones:

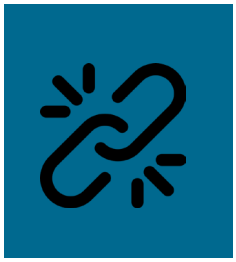
- Ediciones profesionales: Algunos expertos en Photoshop pueden imitar luces y sombras de manera muy realista.
- Escenas complejas: En ambientes con múltiples fuentes de luz (como una discoteca), es más difícil detectar errores.

Resumen:

El análisis de luces y sombras es como **jugar a ser detective con la física de la luz:** si un objeto o persona no "obedece" las reglas de cómo la luz ilumina el mundo real, es probable que sea falso.

6.2 Herramientas de detección

6.2.1 fakeimagedetector.com



Tipo de herramienta:

Como servicio

Forma de funcionamiento:

Análisis de metadatos y ELA.

Pruebas realizadas:

Aunque en la mayoría de las fotos detectó bien en las pruebas realizadas confundió imágenes reales con imágenes, conseguí falsos positivos (es decir que una imagen real me la califiqué como manipulada) pero no un falso negativo (no conseguí pasar una imagen falsa como real).

6.2.2 Image Tampering Detection using ELA and Metadata Analysis

Uso local y demo como servicio.

Análisis de ELA y metadatos.

En desarrollo activo.

6.2.3 ManTraNet

Uso local

Detección de anomalías en imagen

Desarrollo abandonado hace 6 años

Esta herramienta consta de dos redes neuronales.

1. Extractor de características para manipulación de imágenes:
Es una red que analiza y detecta alteraciones en imágenes (como ediciones o retoques). Identifica diferentes tipos de manipulación y convierte los cambios en cada sección de la imagen en un vector de características de tamaño fijo (datos numéricos que representan las alteraciones).

2. Red de detección de anomalías local:
Es una red que compara las características de una zona pequeña de la imagen con el promedio de características de su área circundante. Su funcionamiento se basa en qué tan diferente es una característica local de la referencia (por ejemplo, si hay una anomalía), no en el valor absoluto de esa característica.

Explicación breve:

- La primera red *detecta y describe* alteraciones en imágenes.
- La segunda red *busca diferencias sospechosas* al comparar zonas locales con su entorno.

6.2.4 FAL Detector

URL proyecto	URL código	URL paper:	Tipo de herramienta:	Forma de funcionamien-	Estado del proyecto:
			Uso local	Análisis fragmentos de imagen y comparación con zonas locales.	Desarrollo abandonado hace 6 años

El FAL Detector detecta deepfakes usando dos técnicas clave, simplificadas:

▪ Analiza "parches" (pequeñas secciones de la imagen):
Divide la imagen en partes pequeñas y busca inconsistencias típicas de los deepfakes, como bordes borrosos, cambios de iluminación raros o texturas artificiales.
Convierte cada sección en un código numérico (vector de características) que representa las alteraciones detectadas.

▪ Compara zonas locales con su entorno:
Para cada sección de la imagen, compara sus características con el promedio de las zonas cercanas.
Si una sección tiene características muy diferentes a su entorno (por ejemplo, una cara con sombras extrañas en un fondo uniforme), la marca como sospechosa (anomalía local).

6.2.4 Implementación de detección de ELA en Java

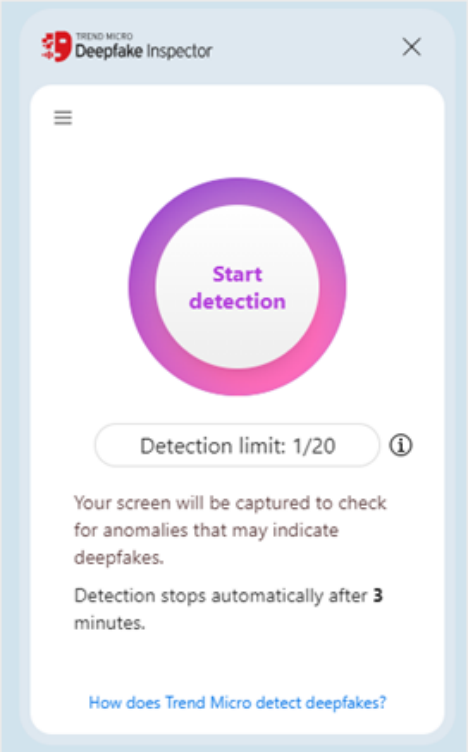
	Tipo de herramienta:	Estado del proyecto:
	Uso local	Abandonado hace 8 años

Detección de manipulación de imagen usando ELA.

6.2.6 Trend Micro Deepfake Inspector

		Tipo de herramienta:	Forma de funcionamiento:	Estado del proyecto:
		Uso local, gratuita	Detección de anomalías en vídeos.	En desarrollo activo.

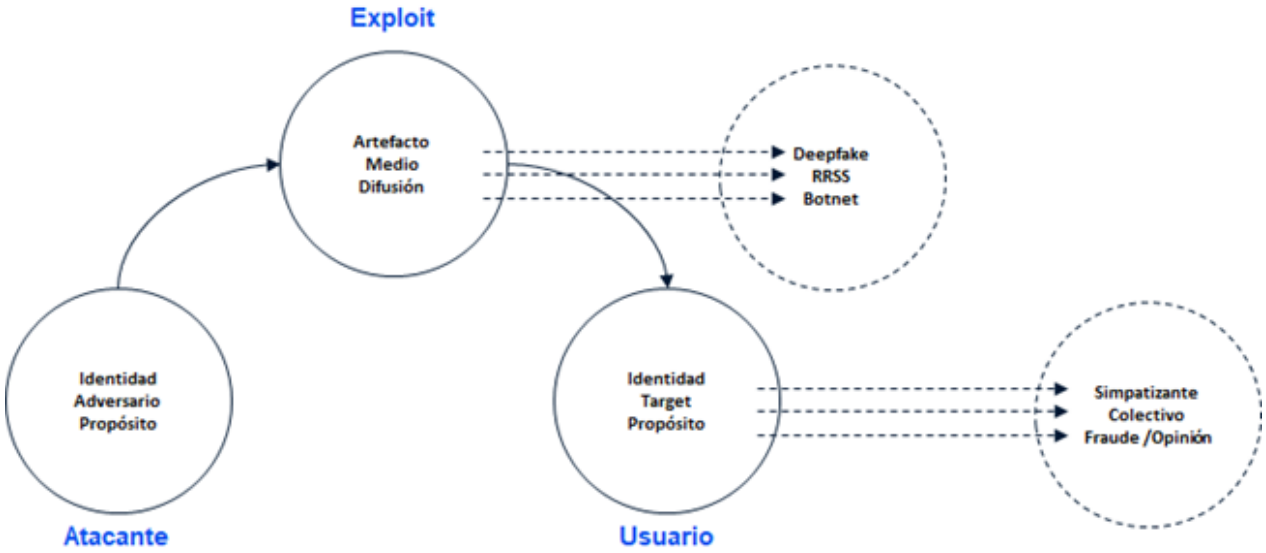
Deepfake Inspector permite no sólo detectar anomalías en vídeos grabados, sino también puede identificar face-swaping en videollamadas en tiempo real con el continuo escaneo que realiza la herramienta. En caso de detección de posible anomalía, alerta al usuario para avisar sobre la posibilidad de estar interactuando con un estafador de deepfake.



7. Ciclo de vida de un Deepfake

En esta sección se intenta expresar que un Deepfake no es sino un tipo de 'artefacto' y como tal un medio para conseguir un fin. Si analizamos el Deepfake como un 'tipo de ataque ciber' debemos pensar en quien fabrica el artefacto, con qué fin y cómo lo distribuye. Asimismo, en el otro extremo hay que valorar quien lo detona y si el daño que produce es personal o colectivo.

El comportamiento de un Deepfake como 'artefacto' es similar al de un ransomware. Si nos abstraemos, adquiere relevancia con 'la viralidad'. La mecánica de distribución sin embargo es el propio usuario y el medio por el que se comparte, en muchos casos 'las redes sociales', aprovechando que éstas promocionan los contenidos llamativos, que se viralizan pronto y que provocan reacciones emocionales del público. Los algoritmos lo llaman 'engagement'. Este elemento hace muy difícil cualquier detección y/o repuesta. De alguna manera, el algoritmo promociona 'la popularidad', en definitiva 'aquello que se convierte en viral'. La única circunstancia por la que la red social podría tener interés real en detener la distribución de un artefacto malicioso sería que a la postre se generase una opinión contraria que claramente le perjudicase o degenerase en perdida de usuarios. Creo que la naturaleza 'cotilla' del ser humano sencillamente favorece que esto no pueda ocurrir. En esta sección y como demostración palpable de esta 'teoría' que requiere de 'observación y seguimiento' así como de soluciones imaginativas en aras de proteger 'la neutralidad de la red'



Dentro del observatorio y hasta aquí se ha puesto foco en el concepto de Deepfake, cómo se produce, casos reales que ponen de manifiesto la incidencia que pueden tener y algunas técnicas que pueden utilizarse en la detección, sin embargo, se nos antojan relevantes otras facetas del problema tales cómo dónde podría tener lugar la detección, cómo actuar ante un escenario que no es determinista, el estado de la regulación y quizás contramedidas potenciales.

Un Deepfake es un 'artefacto' que por su naturaleza se distribuye a través de una 'red social' y que para causar impacto exige cierto grado de 'viralización'.
Hacer notar aquí que:
La red social es el medio a través del que se viraliza el contenido. Sus algoritmos amplifican la difusión de los contenidos más populares y que generan mayor engagement. Sin embargo, las leyes en el momento actual no les responsabilizan 'como distribuidores'. Tan sólo les exige que, en caso de identificar contenido inapropiado, falaz y siempre que este cause un daño deberían actuar retirándolo. Además, las leyes en relación son muy locales, sujetas a interpretación y se encuentran en un proceso de revisión, como en el caso de la UE donde se plantea que los usuarios tendrán que etiquetar los contenidos que se difunden y que han sido generados por IA.

Región	Leyes o Propuestas Clave	Estado	Aspectos Notables
Unión Europea	Ley de Inteligencia Artificial (AIA): Exige identificar contenido generado por IA, como deepfakes.	Aprobada (2024)	Obliga a etiquetar deepfakes como generados por IA. Pone un marco regulatorio más amplio sobre IA.
España	Propuesta de Ley Orgánica sobre Deepfakes: Busca regular deepfakes, especialmente en elecciones y privacidad.	En vigor (2023)	Enfocada en la protección de datos y la privacidad.
Estados Unidos	Texas: Ley sobre deepfakes en elecciones. California: Ley sobre Deepfakes Sexuales. Nueva York: Propuesta para regular el uso de deepfakes en medios. Washington: Regulación general del uso de deepfakes. Illinois: Regulación del uso de deepfake en contextos laborales y educativos.	En vigor/En discusión (varios estados) 2023-2024	Enfatiza en contextos específicos como elecciones y privacidad.
Corea del Sur	Ley Sobre Deepfakes Sexuales: Penaliza la distribución no consensuada.	En vigor (2022)	Incluye multas hasta $22,000 y 3 años de prisión. Abarca aspectos psicológicos, sociales y económicos.
Argentina	Propuesta de Ley S3650-2024: Propone sanciones penales para quienes distribuyan deepfake sin consentimiento.	En discusión (2024)	Busca proteger derechos individuales, enfocándose en el consentimiento.
China	Regulaciones Administración del Ciberespacio: Requiere permisos para plataformas que usen tecnología para crear o manipular datos.	En vigor (2019)	Aplica a redes sociales y plataformas digitales. Auditorías regulares de medios digitales.

Obsérvese, la aproximación de China que entra en conflicto con los derechos y libertades del mundo Occidental y con la neutralidad de la red. El fondo de la cuestión es que a fecha de hoy el editor del Deepfake es el único responsable y si éste lo puede poner en circulación y consigue viralizarlo permaneciendo anónimo no hay forma de perseguirlo.

Corolario: La creación y difusión de deepfakes provocan ciertos conflictos intelectuales entre la libertad de expresión y el daño que un contenido puede causar. Las redes sociales no se consideran responsables por la distribución sin embargo sus algoritmos son el mecanismo que los atacantes utilizan para viralizar o dirigir estos artefactos.

Si son las redes sociales el espacio de difusión y más allá del estado del arte con relación a la tecnología, parece lógico pensar que son las redes sociales donde habría que enfocar la función de DETECCIÓN.

8. Conclusión

La tecnología deepfake ofrece tanto riesgos significativos como oportunidades sustanciales en diversos sectores. El uso ético y responsable puede llevar a avances en marketing, entretenimiento, educación, atención médica y aplicaciones culturales, mejorando las experiencias de los usuarios e impulsando la innovación. Sin embargo, el nivel de riesgo del uso de deepfakes de forma malintencionada crece al mismo ritmo o superior y pudiendo ser su difusión exponencial, por ello no sólo el desarrollo de herramientas que faciliten su detección es necesario, sino que debe ir acompañado de regulación, concienciación y educación pública, esenciales para combatir los posibles daños de los deepfakes.

Los avances en el aprendizaje automático y la inteligencia artificial seguirán mejorando las capacidades del software utilizado para crear deepfakes. Según los expertos, las Redes Generativas Adversarias (GAN), la disponibilidad de conjuntos de datos públicos y el aumento de la potencia de cálculo serán los principales impulsores del desarrollo de deepfakes en el futuro y harán que sean más difíciles de distinguir del contenido auténtico.

Según los diez principales deepfakes ocurridos hasta 2025 seleccionados en el Observatorio, la democratización de herramientas para su desarrollo no solo para la clonación de voz sino también para la suplantación con vídeos en tiempo real, evidencian el uso de diferentes técnicas según su objetivo final.

Para ataques dirigidos con foco en objetivos de fraude financiero, como el fraude del CEO, donde es necesario lograr una total credibilidad de las personas a las que se realiza la estafa, el uso de los deepfakes mediante videos en tiempo real es el método más utilizado. Como cualquier otro ataque de ciberseguridad dirigido, esto conlleva un proceso de ingeniería social con un análisis previo de contexto de la empresa a atacar, recopilación de información de los diferentes interlocutores, incluyendo en este caso imágenes de cada uno de ellos y sus voceas para poder simular una reunión en tiempo real.

En el caso de deepfakes para realizar estafas masivas, generar desinformación o daño moral y reputacional, se emplean métodos más sencillos como suplantación simplemente de voz o imágenes o directamente generación de imágenes con IA a partir de modelos previamente entrenados con imágenes de un contenido específico, ejemplos: Taylor Swift y pornografía infantil.

En los meses y años venideros, según el análisis del observatorio, es muy probable que los actores malintencionados hagan un uso cada vez mayor de la tecnología de deepfake para facilitar diversos actos criminales, con mayor foco en campañas de desinformación para influir o distorsionar la opinión pública, por su mayor facilidad tanto a la hora de crear el artefacto de deepfake como de viralizar su contenido a través de medios como redes sociales. A pesar de ser los algoritmos de las redes sociales los que amplifican la difusión de los contenidos más populares, las leyes en el momento actual no les responsabilizan 'como distribuidores', siendo el editor del deepfake el único responsable.

La creciente disponibilidad de desinformación y deepfakes tendrá un impacto profundo en la forma en que las personas perciben la autoridad y los medios de información. Los expertos temen que esto pueda llevar a una situación en la que los ciudadanos ya no tengan una realidad compartida, o podría crear confusión social sobre qué fuentes de información son fiables; una situación a veces referida como 'apocalipsis de la información' o 'apatía de la realidad'.

Por su parte, el uso de técnicas y tecnologías para la detección de deepfakes también afronta grandes retos, ya que a medida que mejoran las capacidades de detección, también lo hace la calidad de los deepfakes, con los diferentes modelos de entrenamiento aprendiendo continuamente a producir falsificaciones más convincentes. En el documento se han descrito algunos ejemplos de herramientas de software libre para dar soporte a la detección, entre las que se encuentra Deepfake Inspector de Trend Micro, si bien es cierto que requieren de un continuo desarrollo para mantenerse actualizado y evolucionar sus técnicas, entre las que deben incluir mecanismos para rastrear y verificar la autenticidad del contenido. El apoyo tanto de fabricantes específicos de tecnologías de seguridad, como nuevos actores más especializados en casos de uso de biometría y detección de fraude, se hace esencial para combatir desde todos los ámbitos de la sociedad los riesgos que supone esta amenaza creciente
.
El aumento en el uso de deepfakes requerirá legislación para establecer directrices y hacer cumplir la normativa, la cual según se ha mostrado es de ámbito local y en el caso de la UE por el momento, solo se plantea que los usuarios etiqueten los contenidos que se difunden y que han sido generados por IA. Además, las redes sociales y otros proveedores de servicios en línea deberían desempeñar un papel más importante en la identificación y eliminación de contenido deepfake de sus plataformas.

Para abordar los desafíos que presentan los deepfakes, en la UE existen diversas políticas e intentos regulatorios para hacerlos frente. Esto debe ir acompañado del desarrollo de tecnologías que faciliten tanto a las fuerzas del orden como a los ciudadanos, también en su rol de empleados, la detección deepfakes, siendo todavía en la actualidad la concienciación y educación pública las claves para reducir el riesgo de manipulación y desinformación, y a su vez las bases para que la tecnología deepfake se utilice de manera ética y beneficiosa, salvaguardando la confianza pública y promoviendo la innovación responsable.

9. Referencias

- https://www.ucl.ac.uk/news/2020/aug/deepfakes-ranked-most-serious-ai-crime-threat
- Back to the Hype: An Update on How Cybercriminals Are Using GenAI - Noticias de seguridad | Trend Micro (ES)
- https://istitlaa.ncc.gov.sa/en/transportation/ndmo/deepfakesguidelines/Documents/SDAIA_Deepfakes%20Guidelines.pdf
- https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2023/deepfakes-real-threat.pdf
- https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf
- https://veridas.com/
- https://helpcenter.trendmicro.com/en-us/article/tmka-20062
- https://news.trendmicro.com/2024/07/29/deepfakes-101/
- https://www.europol.europa.eu/media-press/newsroom/news
- https://edition.cnn.com/2024/01/25/tech/taylor-swift-ai-generated-images/index.html
- https://www.newtral.es/tiktok-estafa-leonor-te-lo-explicamos/20241209/
- https://www.ucl.ac.uk/news/2020/aug/deepfakes-ranked-most-serious-ai-crime-threat

DEEPFAKES: RIESGOS, CASOS REALES Y DESAFÍOS EN LA ERA DE LA IA

isms forum GIA | GRUPO DE INTELIGENCIA ARTIFICIAL

@ISMSFORUM