Actualizado: marzo 2022

WITNESS ayuda a las personas a utilizar el video y la tecnología para proteger y defender los derechos humanos – witness.org. Para obtener más información sobre nuestro trabajo sobre deepfakes y cómo prepararte mejor, consulta: wit.to/Synthetic-Media-Deepfakes

Deepfakes

Los *deepfakes* facilitan la manipulación o falsificación de voces, rostros y acciones de personas reales, así como la capacidad de afirmar que cualquier video o audio es falso. Se han convertido en una preocupación crítica para celebridades y politícxs, y para muchas mujeres en todo el mundo. A medida que se vuelven más fáciles de hacer, WITNESS aboga por respuestas a los daños actuales y por la preparación para las amenazas futuras que afectan a las poblaciones vulnerables a nivel mundial. En este panorama general sobre los Deepfakes exploramos:

- **Tecnología**: ¿Cuáles son las tecnologías clave de *deepfakes* y qué pueden hacer?
- **Amenazas**: ¿Cuáles son las principales amenazas identificadas a nivel mundial?
- **Soluciones**: ¿Cuáles son las posibles soluciones técnicas y de política pública?

¿Qué son los deepfakes y los medios sintéticos?

Los *deepfakes* son nuevas formas de manipulación audiovisual que permiten a las personas crear simulaciones realistas de la cara, la voz o las acciones de alguien. Permiten a las personas hacer que parezca que alguien dijo o hizo algo que no hizo o que sucedió un evento que nunca aconteció. Son cada vez más fáciles de hacer, requieren menos imágenes de origen para construirlos y las herramientas para crearlas se comercializan cada vez más. Actualmente, los *deepfakes* tienen un impacto desproporcionado en las mujeres porque se utilizan para crear imágenes y videos sexuales no consensuados con el rostro de una persona específica. Pero existe el temor de que los *deepfakes* tengan un impacto más amplio en la sociedad, empresas y en la política, así como en las investigaciones de derechos humanos, la recopilación de noticias y los procesos de verificación.

Los *deepfakes* son solo un desarrollo dentro de una familia de técnicas habilitadas por inteligencia artificial (IA) para la generación de medios sintéticos. Este conjunto de herramientas y técnicas permiten la creación de representaciones realistas de personas que hacen o dicen cosas que nunca hicieron, la creación realista de personas y objetos que nunca existieron o de eventos que nunca sucedieron.

La tecnología de los medios sintéticos actualmente permite estas formas de manipulación:

- **Añadir y eliminar objetos de un video con mayor facilidad**

- **Modificar las condiciones del fondo de un video.** Por ejemplo, cambiar el clima para que un video grabado en verano parezca que se grabó en invierno
- **Fingir movimientos faciales o corporales ("tiritar"):** Simular y controlar una representación de video realista de los labios, las expresiones faciales o el movimiento del cuerpo de una persona específica (por ejemplo, para hacer creer que estaba borracha).
- **Sincronización labial falsa:** Hacer coincidir una pista de audio con una manipulación realista de los labios de alguien para que parezca que ha dicho algo que nunca hizo
- **Voz falsa**: Generar una simulación realista de la voz de una persona específica
- **Cambiar el género de una voz o hacer que suene como alguien más**: Modificar una voz existente con una "piel de voz" de un género diferente, o de una persona específica.
- **Crear una foto realista pero totalmente falsa de una persona que no existe.** La misma técnica también puede aplicarse de forma menos problemática para crear falsas hamburguesas, gatos, etc.
- **Crear una foto de un evento u objeto a partir de una descripción de texto**
- **Transferir un rostro realista de una persona a otra, la forma más conocida de "*deepfake*"**

[Observa ejemplos de muchos de estos aquí]

¿Cómo funciona la tecnología detrás de ellos?

Estas técnicas principalmente *pero no exclusivamente* descansan en una forma de inteligencia artificial conocida como aprendizaje profundo y del trabajo de Redes Generativas Antagónicas, (Generative Adversarial Networks o GANs por sus siglas en inglés).

Para generar un elemento de contenido de medios sintéticos, se comienza recopilando imágenes o videos de origen de la persona o el elemento que se desea falsificar. Una GAN desarrolla la falsificación mediante el uso de dos redes. Una red genera recreaciones plausibles de las imágenes de origen, mientras que la segunda red trabaja para detectar estas falsificaciones. Estos datos de detección se devuelven a la red dedicada a la creación de falsificaciones, lo que le permite mejorar y crear una versión falsa cada vez mejor y mejor de la fuente, por ejemplo, la cara de la persona a la que se está imitando.

A principios de 2022, muchas de estas técnicas -en particular la creación de *deepfakes* de intercambio de rostros realistas- continúan requiriendo una potencia computacional significativa, una comprensión de cómo ajustar su modelo y, a menudo, una post-producción CGI (Imágenes Generadas por Computadora, o CGI por sus siglas en inglés) significativa para mejorar el resultado final. Buenos ejemplos de un *deepfake* sofisticado que requiere todas estas entradas son los ¡TikTok de“Tom Cruise” videos que puede ser que ya hayas visto!

Sin embargo, incluso con las limitaciones actuales, las personas ya están siendo engañadas por los medios simulados. Como ejemplo, una investigación demostró que las personas no podían detectar de manera confiable las formas actuales de labios, la modificación del movimiento, que se utilizan para hacer coincidir la boca de alguien con una nueva pista de audio. Una investigación reciente

encontró que las personas no son capaces de detectar rostros realistas de personas que nunca existieron. No debemos asumir que las personas están inherentemente equipadas para detectar la manipulación de los medios sintéticos.

El panorama actual de los deepfakes y los medios sintéticos

Los *deepfakes* maliciosos y los medios sintéticos - aún- no están muy extendidos fuera de las imágenes sexuales no consensuadas. Desafortunadamente, los *deepfakes* sexuales no consensuados están disponibles fácilmente y se generan involucrando a celebridades, actrices de la pornografía o a personas comunes.

Además

- Las personas han comenzado a desafiar el contenido real, descartándolo como *deepfake*.
- Aunque la sátira *deepfake* ofrece nuevas oportunidades para la libertad de expresión, a menudo se encuentra en la fina línea del engaño.
- Las imágenes de "personas que nunca existieron" se utilizan cada vez más para disfrazar cuentas falsas en la desinformación.

Las amenazas de los deepfakes

En los talleres liderados por WITNESS así como en las formaciones impartidas a más de 500 personas en los últimos tres años, revisamos los posibles vectores de amenaza con una serie de participantes de la sociedad civil, incluidos los medios de comunicación de base, lxs periodistas profesionales y verificadorxs de datos, con investigadorxs de información errónea y desinformación, así como con especialistas en OSINT (inteligencia de código abierto, u OSINT por sus siglas en inglés). Se priorizaron las áreas en las que las nuevas formas de manipulación podrían ampliar las amenazas existentes, introducir nuevas amenazas, alterar las amenazas existentes o reforzar otras amenazas. También destacaron los desafíos en torno a "es *deepfake*" como primo retórico de "es una noticia falsa".

Lxs participantes en nuestros encuentros de personas expertas en Brasil, África Subsahariana y en el Sudeste de Asia, así como en otras reuniones a nivel mundial, priorizaron sus principales preocupaciones en relación a cómo las nuevas formas de manipulación de los medios, y el aumento de la información errónea y desinformación afectarán su trabajo, sus sociedades y sus comunidades.

- **Periodistas, líderes en la comunidad, y activistas cívicos verán atacadas su reputación y credibilidad,** basándose en las formas existentes de acoso y violencia en línea que se dirigen predominantemente a las mujeres y a las minorías. Ya se han producido varios ataques con

videos modificados contra mujeres periodistas, como en el caso de la destacada periodista india Rana Ayyub.

- **Las figuras públicas se enfrentarán a las imágenes sexuales no consentidas y a la violencia basada en el género, así como a otros usos de los llamados *doppelgangers (dobles)* creíbles.** Lxs políticxs locales pueden ser especialmente vulnerables, ya que cuentan con muchas imágenes pero con menos estructura institucional a su alrededor que lxs políticxs a nivel nacional para ayudar a defenderse de un ataque mediático sintético.
- **Socavar las posibilidades de utilizar el video como evidencia** de los abusos y delitos contra los derechos humanos, obstaculizando la rendición de cuentas y la justicia.
- **Muchxs periodistas, ya sobrecargadxs y con escasos recursos,** no dispondrán de la capacidad forense para analizar medios y verificar hechos en *deepfakes.*
- **Las organizaciones de derechos humanos, de recopilación de noticias y de verificación se verán presionadas para demostrar que algo es cierto, así como para demostrar que algo no fue falsificado.** Los que están en el poder tendrán la oportunidad de utilizar la negación plausible sobre el contenido al poder declarar que es un *deepfake*.
- A medida que los *deepfakes* se vuelven más comunes y más fáciles de hacer en volumen, contribuirán **a la "manguera de la falsedad"** que son estrategias que alimentan a los medios de verificación y a las agencias de verificación de datos con contenidos que tendrán que verificar o desacreditar. Lo que podría sobrecargarlas y distraerlas.
- **Los *deepfakes* intersectarán con los patrones existentes de "incendios digitales" rápidos, donde las imágenes falsas se comparten a gran velocidad y escala a través de** WhatsApp, Telegram y Messenger de Facebook, así como en otras aplicaciones de mensajería.
- **Las videoconferencias en línea** serán vulnerables a la manipulación.

En todos los contextos, las personas a las que consultamos señalaron la importancia de considerar los *deepfakes* en el contexto de los enfoques existentes de comprobación y verificación de datos. Los *deepfakes* y los medios sintéticos se integrarán en las campañas de conspiración y desinformación existentes, a partir de las tácticas (y las respuestas) en evolución en esa área.

La investigación de WITNESS sobre los *deepfakes* y la sátira, incluido un informe reciente titulado Just Joking! identificaron un uso creciente de *deepfakes* para realizar una crítica social y política poderosa. Se demostró cómo se prestan los *deepfakes* satíricos fotorrealistas para:

- **La crítica social:** La parodia y la sátira para criticar al poder que identifican problemas sociales y políticos y que el público reconoce como satíricos.
- **El uso incorrecto deliberado:** Afirmar que algo es una sátira cuando es una desinformación y culpar a las audiencias por no "entender la broma"
- **El mal uso accidental:** Cuando se pierde el contexto, y se comparte como información errónea identificándose como real)

¿Cuáles son las soluciones disponibles?

Existe una gran cantidad de trabajo en curso sobre cómo prepararse mejor para los *deepfakes*. En general, en WITNESS nos preocupa que este trabajo sobre "soluciones" no incluye adecuadamente las voces y las necesidades de las personas que son perjudicadas por los problemas existentes de manipulación de los medios, la violencia del Estado, la violencia basada en el género, y la información errónea en el Sur Global y en las comunidades marginadas del Norte Global.

¿Cómo podemos enseñar a las personas a detectarlos?

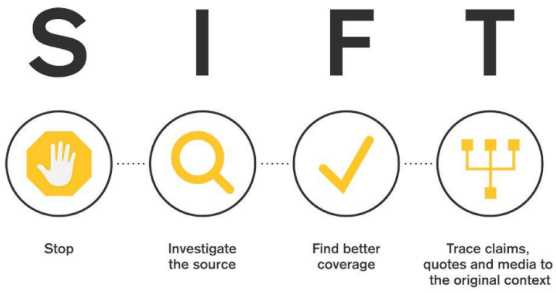
No es una buena idea enseñar a las personas que ellas pueden detectar los *deepfakes* u otras manipulaciones de medios sintéticos. Aunque ahora hay algunos consejos que ayudan a detectarlos - por ejemplo, fallas visibles - estos son solo los errores actuales en el proceso de falsificación y desaparecerán con el tiempo. Sin embargo, si quieres probar tu habilidad puedes visitar: https://detectfakes.media.mit.edu/ (en inglés).

Las plataformas como Facebook, y las empresas independientes desarrollarán herramientas que pueden hacer algo de detección, pero éstas sólo proporcionarán algunas pistas y no estarán ampliamente disponibles en el futuro inmediato. Es importante que las personas también se centren en entender los *deepfaks* dentro de un marco más amplio de alfabetización mediática, como el enfoque SHEEP (en inglés) de la organización First Draft (en inglés) o en el marco SIFT (en inglés) .

SHEEP (un acrónimo en inglés) sugiere que para evitar ser engañadx por la información errónea en línea deberías "pensar en **SHEEP** antes de compartir".

SOURCE (FUENTE): mira lo que hay debajo. Verifica el *about page* (sobre esta página) de un sitio web o una cuenta, mira cualquier información de la cuenta y busca nombres y nombres de usuario.
HISTORIA: ¿la fuente tiene una agenda? Averigua qué temas se cubren con regularidad o si promueve únicamente una perspectiva.
EVIDENCIA: explora los detalles de un reclamo o meme para identificar si está respaldada por evidencia fiable de otros lugares.
EMOCIÓN: ¿la fuente se basa en una emoción para hacer un punto? Busca lenguaje sensacionalista, incendiario y divisivo.
PICTURES (FOTOGRAFÍAS): las imágenes pintan mil palabras. Identifica qué mensaje está retratando una imagen y si la fuente está utilizando imágenes para llamar la atención.

SIFT (por sus siglas en inglés) proporciona otro modelo relacionado con el análisis desde el sentido común de la información sospechosa:



Traducción de la imagen:
Stop: Alto
Investigate the source: Investiga la fuente
Find better coverage: Encuentra una mejor cobertura
Trace claims, quotes and media to the original context: Identifica reclamos, citas y medios al contexto original

¿Cómo construimos sobre la capacidad y la coordinación periodísticas existentes?

Lxs periodistas y lxs investigadorxs de los derechos humanos deben comprender mejor acerca de cómo detectar *deepfake* utilizando las prácticas existentes de OSINT y combinandolas con las nuevas herramientas forenses para multimedia que se están desarrollando. Aprende más en el informe de WITNESS sobre necesidades (en inglés), y el trabajo reciente Partnership on AI (en inglés).

¿Estas herramientas son útiles para la detección? y ¿quién tiene acceso?

La mayoría de las plataformas y muchos *start-ups* están desarrollando herramientas para la detección de *deepfakes*.

Algunas herramientas comienzan a ser publicadas. Un ejemplo es Sensity.AI (en inglés) https://platform.sensity.ai/deepfake-detection

Sin embargo, recomendamos abordarlos con extrema precaución. Los recientes y amplios concursos para la detección de *deepfakes* no han aportado modelos suficientemente eficaces con las técnicas conocidas o suficientemente aplicables a las nuevas técnicas, y la mayoría de los detectores disponibles públicamente serán menos eficaces que los sistemas cerrados. Las herramientas de detección tienden a ser menos confiables si no se conoce la técnica utilizada para generar los medios sintéticos, y menos fiables en los medios de baja resolución o comprimidos que vemos en línea. Una experiencia reciente de un deepfake sospechoso en Myanmar muestra los desafíos de confiar en los detectores disponibles públicamente sin la experiencia necesaria que los acompaña.

Aunque se desarrollen herramientas sólidas, estas no estarán disponibles de forma generalizada, particularmente fuera de las plataformas y las empresas de medios. Es probable que los medios y las organizaciones de la sociedad civil en el Sur Global queden al margen y es importante que aboguen por los mecanismos que les permitan tener un mayor acceso a las instalaciones de detección. En WITNESS abogamos por el incremento de la equidad en el acceso a las herramientas de detección de *deepfakes,* en invertir en las habilidades y en la capacidad de la sociedad civil y periodistas a nivel mundial, y para el desarrollo de mecanismos de escalada que provean de análisis sobre sospechas críticas de *deepfakes*.

¿Existen herramientas de autenticación? y ¿a quién se excluye?

Existe un movimiento creciente para desarrollar herramientas que permitan rastrear mejor la procedencia de los videos y las imágenes, desde el momento en que se graban en los teléfonos inteligentes hasta que se editan, y luego se comparten o distribuyen en las redes sociales. Esta "infraestructura de procedencia y autenticidad" puede mostrarte información sobre la procedencia de una fotografía o video y sobre si se ha modificado /y cómo una fotografía. Esto es relevante tanto para la manipulación de video "superficial" (o "*shallowfake*" en inglés) como para los videos mal contextualizados o editados, así como para los *deepfakes*. Puedes después utilizar esta información para ayudarte a tomar decisiones sobre si confiar en el contenido. Un ejemplo de una iniciativa en esta área es la Iniciativa de autenticidad de contenido (en inglés) dirigida por Adobe, y la

recientemente lanzada Coalición para la Autenticación y Procedencia del Contenido (C2PA por sus siglas en inglés).

Sin embargo, existe el riesgo de que las herramientas que se desarrollen para ayudar a rastrear mejor los orígenes de los videos y mostrar cómo han sido manipulados pueden también crear riesgos de vigilancia y exclusión para las personas que no quieran añadir datos e información adicionales a sus fotografías o videos, o que no pueden atribuirse las fotografías por miedo a lo que los gobiernos y las empresas harán con la información. WITNESS ha liderado *Harm, Misuse and Abuse Assessment* (análisis de daños, uso indebido y abuso) de las especificaciones del C2PA para identificar estos y otros daños potenciales, y para desarrollar estrategias para evitarlos y mitigarlos. Esta "infraestructura de procedencia y autenticidad" seguirá siendo objeto de abusos, y los actores malintencionadxs encontrarán lagunas, por lo que el paso clave para avanzar es reforzar un marco de derechos, con barreras contra los daños, mecanismos de reparación, y con oportunidades para empoderar las voces críticas.

¿Qué deberían de hacer las plataformas?

Las plataformas de redes sociales como Facebook y Twitter tienen políticas sobre *deepfakes* y acerca de cómo los manejan, y sobre cómo manejan los medios manipulados de manera más amplia.

WITNESS analiza la política de Facebook aquí y la política de Twitter aquí

Los elementos clave de estas políticas incluyen:

- ¿Cubren sólo los *deepfakes* o también otras formas de medios manipulados (por ejemplo, ¿un video ralentizado, o un video mal contextualizado?)
- ¿Cómo se define el daño causado por un video?
- ¿Importa la intención de compartir?
- ¿Retiran un video ofensivo? ¿Lo etiquetan? ¿Proporcionan un contexto sobre la manipulación? ¿Lo hacen menos visible en su sitio o menos fácil de compartir?
- ¿Se aplican a figuras públicas?

Facebook (Meta)

La política de Facebook es específica para los *deepfakes* en lugar de otras formas de manipulación de videos o fotos.

Facebook removerá medios manipulados cuando

- "Fue editado o sintetizado, más allá de ajustes para mejorar su claridad o calidad, de manera que no son aparentes para una persona común y que podría llevar a pensar que alguien dijo palabras que en realidad no pronunció.

- Es producto de la inteligencia artificial o del aprendizaje automático que combina, reemplaza o superpone contenido en un video, haciendo que parezca auténtico.

Esta política no se extiende al contenido que sean parodia o sátira, ni a los videos que hayan sido editados únicamente para omitir o cambiar el orden de las palabras. Otros videos manipulados y engañosos pueden ser referidos o recogidos por sus verificadorxs de datos de terceros. Existen ejemplos en los que Facebook ha retirado erróneamente la sátira *deepfake* como información errónea. Un ejemplo de esto ocurrió en Camerún cuando un académico y activista local compartió un video claramente fabricado del Embajador de Francia diciendo a lxs cameruneses que nunca lograron realmente la independencia de la explotación colonial de Francia. Lxs verificadorxs de datos de terceros de Facebook en la emisora francesa France 24 calificaron el video como parcialmente falso, nulificando así el poder retórico de la crítica.

El audio, las fotografías o los videos, ya sea un *deepfake* o no, será eliminado de Facebook si viola nuestras Normas Comunitarias, incluyendo aquellas que regulan la desnudez, la violencia gráfica, contenido que procura suprimir votos y discurso de odio."

Twitter
La política de Twitter está disponible aquí

Twitter indica que "no puedes compartir elementos multimedia sintéticos, alterados o sacados de contexto (no sólo *deepfakes* sino también otras formas de manipulación) que puedan engañar o confundir a las personas y provocar daños . Asimismo, podemos etiquetar los Tweets que incluyen contenido multimedia engañoso para que la gente tenga información sobre su autenticidad y para ofrecer más contexto."

Twitter se enfoca en tres preguntas clave que determinan si podrían o si etiquetarán el contenido o lo removerán.

1. ¿El contenido es sintético o está manipulado?
2. ¿El contenido se comparte de manera engañosa?
3. ¿El contenido podría tener un impacto en la seguridad pública o provocar daños graves?

¿Se altera el contenido de forma significativa y engañosa o se fabrica?	¿El contenido se comparte de forma engañosa?	¿Es probable que el contenido afecte a la seguridad pública o cause un daño grave?	
✓	✗	✗	El contenido **puede** ser etiquetado.
✗	✓	✗	El contenido **puede** ser etiquetado.
✓	✗	✓	El contenido **podría** ser etiquetado, o **puede** ser eliminado.
✓	✓	✗	El contenido es **probable** que sea etiquetado.
✓	✓	✓	El contenido es **probable** que sea eliminado.

TikTok

TikTok "prohíbe falsificaciones digitales (completa o parcialmente manipuladas) que engañen a los usuarios distorsionando la verdad de los hechos y que causen daños sustanciales al sujeto del vídeo, a otras personas o a la sociedad."

Nuestra opinión

Las plataformas deben ser proactivas a la hora de señalar, rebajar el rango -y en el peor de los casos, eliminar- los *deepfakes* maliciosos, porque lxs usuarixs tienen poca experiencia con este tipo de manipulación invisible para los ojos e inaudible para los oídos, y porque lxs periodistas no tienen las herramientas necesarias para detectarlos de forma rápida o eficaz. Pero abordar los *deepfakes* no elimina la responsabilidad de abordar también activamente otras formas de manipulación de videos "superficiales", como etiquetar erróneamente un video real o editarlo ligeramente.

Algunas preguntas en curso que se relacionan con las políticas:

- ¿Cómo garantizarán tanto Facebook, Twitter, TikTok y otros la detección precisa de deepfakes?
- ¿Cómo harán las plataformas para juzgar cuándo una modificación es maliciosa, o si algo es parodia, o por el contrario se hace pasar por sátira o parodia?
- ¿Cómo comunicarán lo que aprendan a lxs consumidorxs escépticxs?
- ¿Cómo se asegurarán de que cualquier decisión que tomen esté sujeta a la transparencia y apelación, ya que cometerán errores?

¿Qué están haciendo las personas legisladoras?

Los gobiernos están empezando a legislar sobre los *deepfakes*. Estas leyes se refieren tanto a las imágenes sexuales no consensuales como a los usos para el engaño y la información errónea.

En los Estados Unidos de Norteamérica se han propuesto varias leyes a nivel estatal y federal, y en la Unión Europea, la ley de IA avala el etiquetado de los medios sintéticos para la protección de las personas consumidoras.

En la región de Asia-Pacífico, existen dos ejemplos son las leyes de la República Popular China, que prohíben *deepfakes* y otras "noticias falsas", y la legislación recientemente propuesta en Filipinas. Hay que tener cuidado con estas leyes cuando hacen una definición muy amplia de la falsificación audiovisual e incluyen formas importantes de libertad de expresión como la sátira, o dan amplia discreción y poder a los gobiernos para decidir lo que es "falso".