Deepfake, cuando lo que vemos ya no es de fiar

La vista es uno de los sentidos en los que más confiamos para entender la realidad. De alguna forma, creemos más en lo que vemos con nuestros propios ojos, que en lo que nos cuentan o leemos. Pero, ¿qué pasa cuando ya no tenemos la certeza de que lo que contemplamos es real? Las *deepfake* son falsificaciones de imágenes creadas por inteligencia artificial que abren la terrible duda sobre si nos debemos creer todo lo que vemos.

> "¿Por qué Stalin borró a esas personas de esas fotografías? ¿Por qué se tomó la molestia? Es porque hay algo muy, muy poderoso relacionado con la imagen visual. Si cambias la imagen, cambias la historia. Somos seres increíblemente visuales. Confiamos en la visión – e, históricamente, ha sido muy fiable. Y por ello, las fotos y los vídeos todavía tienen esta increíble resonancia[1]".

Son las palabras de Hany Farid, experto forense en fotografía digital, recordando cómo el dirigente soviético no dudaba en retocar las imágenes en las que aparecían sus oponentes políticos –como lo fue León Trotsky-, eliminando así su presencia de la Historia.

"Si no lo veo, no lo creo", decimos con frecuencia, otorgando a la imagen un estatus de veracidad de lo que muestra, que hoy en día podría estar en peligro. El propio Farid se pregunta en el artículo hasta cuándo podremos seguir confiando sin dudar en lo que vemos en fotografías y vídeos. Y es que la posibilidad que nos otorga la tecnología de falsear los contenidos audiovisuales, pone en cuestión la realidad tal y como la conocemos.

No es un problema que afecte solamente a la interpretación de los hechos que pueda tener lugar en un juicio, por poner un ejemplo práctico de situaciones en las que modificar las pruebas es un tema gravísimo; en el periodismo, la evidencia visual es una piedra angular en la creación de la opinión pública, y, además, es un elemento fundamental para determinar cómo se conforma el poder político. De alguna forma, las manipulaciones de archivos audiovisuales –tanto de vídeo, como de audio- suponen un peligro para la convivencia democrática y para la dignidad de las personas afectadas, que ven cómo los registros de su imagen son alterados con el fin de desacreditarlas, o de convertirlas en portadoras de un mensaje u opinión con el que no comulgan y al que no apoyan.

Hablamos del *deepfake*, un fenómeno que hoy en día afecta especialmente al vídeo, dada la capacidad que ha desarrollado la inteligencia artificial para trucar, con un grado de éxito más que notable, cualquier pieza audiovisual, poniendo en boca de políticos afirmaciones que no han realizado, o –y esto es lo más común- alterando escenas de películas pornográficas situando los rostros de personajes conocidos en el cuerpo de los actores enfrascados en actos sexuales.

De acuerdo con BBC News, en los últimos nueve meses se ha duplicado el número de vídeos falsos que proliferan por las redes. La empresa tecnológica Deeptrace ha llegado a detectar más 14 600, frente a los menos de 8 000 encontrados en diciembre de 2018. De ellos, el 96% eran de carácter pornográfico, generalmente con la cara de una celebridad generada por ordenador sobre el cuerpo de un actor o de una actriz de la trama porno. Por cierto, que el *deepfake*, aparte de una herramienta para condicionar la opinión pública, supone un lucrativo negocio para algunos.

> En los últimos nueve meses se ha duplicado el número de vídeos falsos que proliferan por las redes.

Según expone el informe de Deeptrace, los cuatro principales sitios web que albergan este tipo de vídeos pornográficos trucados atrajeron a 134 millones de usuarios desde febrero de 2018. Calcule usted lo que supone eso en ingresos por la publicidad incluida en esos portales.

El mal uso del audiovisual sintético

La palabra *deepfake* procede de la contracción de *deep learning* (aprendizaje profundo) y *fake* (falsificación). Es decir, que implica el uso de inteligencia artificial para generar vídeos sintéticos, generalmente con el fin de desacreditar a alguien y/o condicionar la opinión pública. Hace algún tiempo apareció en la red Instagram un vídeo de Mark Zuckerberg, el popular consejero delegado de Facebook, en el que este confesaba su intención de hacerse con el control del planeta, gracias a disponer de los datos de las personas. Incluso hacía un guiño al cine de James Bond al mencionar a la organización Spectra, la archienemiga del agente 007. Esto es un ejemplo de lo que se puede hacer en el campo del *deepfake*.

> Hace algún tiempo apareció en la red Instagram un vídeo de Mark Zuckerberg en el que confesaba su intención de hacerse con el control del planeta, gracias a disponer de los datos de las personas.

En cualquier caso, se trata de un fenómeno relativamente reciente. El primer caso conocido de la manipulación facial a través del uso de inteligencia artificial tuvo lugar hace apenas dos años, cuando un usuario de Reddit subió a la red una serie de vídeos pornográficos, en los que aparecían los rostros de conocidas actrices, como el de Gal Gadot o el de Scarlett Johansson. Curiosamente, el nombre del usuario era precisamente Deepfake.

La organización Witness introduce el *deepfake* dentro del marco conceptual del desorden informativo. Resulta especialmente preocupante el impacto en las personas de la información visual, bastante más fuerte que la textual, dado que nuestros cerebros tienden a confiar más en la imagen. El análisis de esta oenegé distingue tres aspectos distintos: *misinformation* (misinformación), cuando la mala información no ha sido producto de mala intención, sino de un error o equívoco; *malinformation* (malinformación), el difundir información verdadera, pero de carácter privado, con la intención de hacer daño (por ejemplo, airear un vídeo íntimo de alguien manteniendo relaciones sexuales); y, entre ambas, la *disinformation* (desinformación), que implica

crear y difundir información falsa con malas intenciones. Las *deepfake* entrarían dentro de esta categoría.

Fuente: Witness (2018) “Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening”.

Existen numerosas herramientas en el mercado para construir vídeos y audios falsos. Witness realiza la siguiente clasificación de las mismas:

- *Audio simulado individualizado*: se trata de software capaz de imitar la forma de hablar de una persona, como Lyrebird o Baidu DeepVoice.
- *Editores que permiten cambiar elementos centrales o del fondo de la imagen*: en este apartado mencionan Adobe Cloak.
- *Recreación facial*: se utilizan para manipular el rostro de una persona en un vídeo. Productos como Face2Face y Deep Video Portraits permiten transferir los movimientos de la cara y el busto de alguien a otra persona.
- *Reconstrucción facial realista y sincronización del movimiento de los labios sobre un audio existente.*
- *Personajes reales a las que se les ha cambiado una parte del cuerpo, en general, el rostro*: se puede llevar a cabo con herramientas como FakeApp o FaceSwap.

La magnitud de la mentira

Desde su aparición en 2017, el fenómeno *deepfake* ha ido cobrando volumen en muy poco tiempo, habiendo crecido el número de casos detectados a pasos agigantados. Un informe de la empresa holandesa Deeptrace establece un incremento del 100%, respecto del año pasado; de 7 964 casos que detectaron en diciembre de 2018, hasta los más de 14 600 registrados en septiembre de este año.

La mayor parte de estos vídeos –el 96%- es de tipo pornográfico. El éxito de este tipo de contenido es innegable: a pesar del escaso tiempo que llevan proliferando por las redes, los cuatro principales portales web que acogen este tipo de obras de porno falseado han superado los 134 millones de visualizaciones.

La mayor parte de estos vídeos –el 96%- es de tipo pornográfico.

Otro dato interesante es que el *deepfake* de naturaleza pornográfica se centra solamente en hacer daño a mujeres, mientras que, en los otros tipos de vídeos, los afectados son en una ligera mayoría varones (en un 61% de los casos).

Por otro lado, en las falsificaciones no pornográficas los protagonistas son personalidades mayormente occidentales (de Estados Unidos, Reino Unido y Canadá), pero en los de contenido sexual, el peso de Asia aumenta, en concreto, procedente de Corea de Sur, que concentra una cuarta parte de los casos, principalmente en torno a los cantantes pop.

Si nos fijamos en la profesión de las víctimas de *deepfakes*, en el caso de las pornográficas, las afectadas son casi en su totalidad mujeres relacionadas con la industria del entretenimiento –actrices y cantantes-, aunque

en los audiovisuales sin contenido sexual, si bien mayoritariamente están dirigidos a personajes del espectáculo, hay también una presencia de políticos y de profesionales de los medios de comunicación.

DeepNude: la visión de rayos x

Uno de los ejemplos de herramientas de falsificación de imágenes que más ha dado que hablar en los últimos tiempos ha sido DeepNude, que no es otra cosa que un algoritmo que permite "desnudar" a las mujeres en las fotografías (parece ser que con los varones no funciona).

Se trata de una web, lanzada en junio de 2019 por un estonio que dice llamarse Alberto, y que en poco tiempo ha tenido un volumen de visitas record. A través del portal en cuestión o de una app para móvil, el usuario puede subir una fotografía, y el algoritmo en cuestión (pix2pix, desarrollado por la Universidad de California en 2017 y entrenado con más de 10 000 fotografías de desnudos), se encarga de reproducir las partes del cuerpo no visibles, incluyendo los órganos sexuales femeninos.

> DeepNude es un algoritmo que permite "desnudar" a las mujeres en las fotografías.

A pesar de que tanto la versión gratuita como la de pago de la aplicación imprimen una marca de agua en la foto resultante, avisando de su falsedad, no es poca la preocupación que despiertan inventos como este. Teniendo en cuenta lo fácil que resulta encontrar imágenes de desnudos de mujeres en internet, no parece que esa sea la principal utilidad que un usuario persiga al utilizarla. En cambio, parece una herramienta que puede ser fácilmente utilizada para hacer daño, pues a pesar de que pronto se detecte la falsedad de una de estas fotos, el perjuicio público que le puede infligir a una víctima es irreversible.

Redes Generativas Antagónicas

Muy relacionadas con el *deepfake* están las redes generativas antagónicas (en inglés *Generative Adversarial Networks* o GANs), un tipo de redes neuronales capaces de generar un elemento falso en un vídeo, como, por ejemplo, un rostro conocido sobre el de una actriz porno, una práctica muy en boga, como hemos podido comprobar.

Básicamente, se trata de una red neuronal –inteligencia artificial basada en el aprendizaje profundo- que es enfrentada con otra. La primera es conocida como el generador y la segunda como discriminador. El funcionamiento es el siguiente: el generador crea muestras falsas (una imagen, un vídeo o un texto, por ejemplo) intentando engañar al discriminador, haciéndole creer que son reales. Este último a su vez, debe determinar si la creación es verdadera, obligando al primero a superarse en destreza cada vez. Al acabar este entrenamiento, habremos conseguido un generador realmente bueno en producir contenidos falsos difíciles de distinguir de los reales.

Las GAN se convierten en una herramienta muy potente a la hora de crear imágenes y audiovisuales falsos, pero que presentan un grado de perfección tal que resultan muy difíciles de distinguir de algo verdadero.

La batalla contra el *deepfake*

La preocupación ante el daño que puede producir este delito emergente, tanto a la reputación de personas concretas como por la posibilidad de manipulación de la opinión pública, ha llevado a que se estén tomando medidas para frenarlo en la medida de lo posible.

El Gobierno de los Estados Unidos ha puesto en marcha el proyecto *Media Forensics (MediFor)* desde DARPA (*Defense Advanced Research Projects Agency*) para la investigación y desarrollo de tecnologías capaces de certificar la autenticidad de una imagen o vídeo, y de detectar automáticamente las manipulaciones.

No obstante, también existen iniciativas en este sentido desde el sector privado, en concreto desde Google y Facebook.

Google ha puesto a disposición del público una base de datos con más de 3 000 vídeos *deepfake* en los que ha sido utilizada inteligencia artificial para alterar los rostros de las personas que aparecen. El objeto es apoyar a los investigadores en este campo en el desarrollo de herramientas contra este delito, mediante los conocimientos que puedan adquirir de este gran banco de datos.

Google ha puesto a disposición del público una base de datos con más de 3 000 vídeos deepfake.

Por su parte, la empresa de Zuckerberg ha creado el proyecto *Deepfake Detection Challenge* (DFDC), en colaboración con socios como el MIT, Microsoft y varias universidades estadounidenses. La iniciativa persigue el desarrollo de tecnología que todo el mundo pueda utilizar para poder saber cuándo un vídeo ha sido manipulado. En concreto, se prevé la creación de una gran base de datos y de un programa de ayudas y becas para la investigación.

Photo by **Mike** from **Pexels**

BBC News (2019) "Google makes deepfakes to fight deepfakes". Disponible en: https://www.bbc.com/news/technology-49837927

Cellan-Jones, R. (2019) "Deepfake videos 'double in nine months'" en *BBC News*. Disponible en: https://www.bbc.com/news/technology-49961089

Deeptrace (2019) "The State of Deepfakes. Landscape, Threats and Impact". Disponible en: https://storage.googleapis.com/deeptrace-public/Deeptrace-the-State-of-Deepfakes-2019.pdf

Goodfellow, J., Pouget-Abadie, J. Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. y Bengio, Y. (2014) "Generative Adversarial Nets". Université de Montréal. Disponible en: https://arxiv.org/pdf/1406.2661.pdf

Pérez, E. (2019) "DeepNude: la polémica aplicación que «desnuda» a cualquier mujer mediante inteligencia

artificial" en *Xataca*. Disponible en: https://www.xataka.com/privacidad/deepnude-polemica-aplicacion-que-desnuda-a-cualquier-mujer-mediante-inteligencia-artificial

Schroepfer, M. (2019) "Creating a data set and a challenge for deepfakes" en *Facebook Artificial Intelligence*. Disponible en: https://ai.facebook.com/blog/deepfake-detection-challenge/

Taulli, T. (2019) "Deepfake: What You Need To Know" en *Forbes*. Disponible en: https://www.forbes.com/sites/tomtaulli/2019/06/15/deepfake-what-you-need-to-know/#3de32a52704d

Witness (2018) "Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening".

Yvas, K. (2019) "Generative Adversarial Networks: The Tech Behind DeepFake and FaceApp" en *Interesting Engineering*. Disponible en: https://interestingengineering.com/generative-adversarial-networks-the-tech-behind-deepfake-and-faceapp