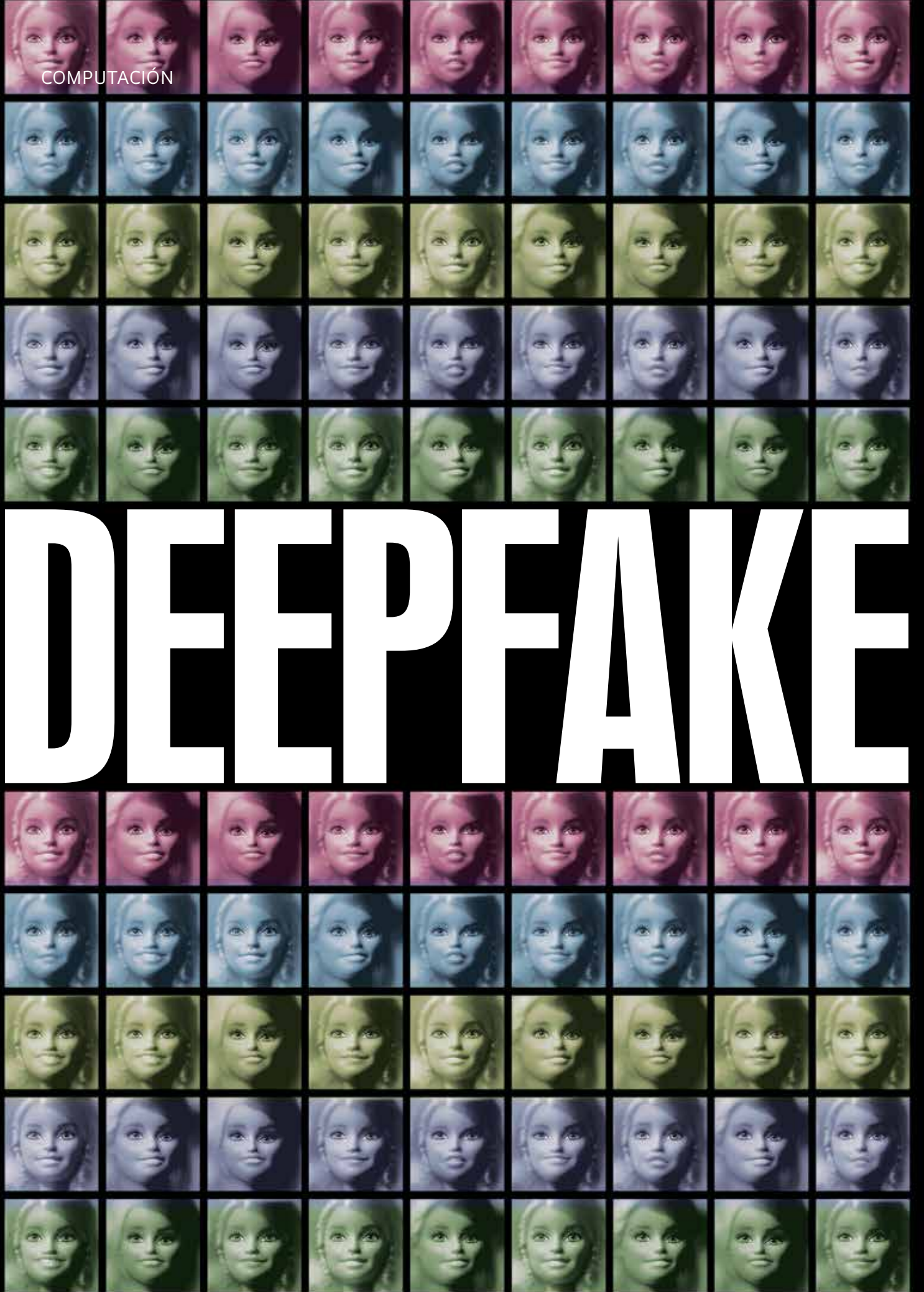

DEEPFAKE

Un algoritmo detecta imágenes y videos adulterados mediante inteligencia artificial, el nuevo nivel del desarrollo tecnológico para propagar la desinformación

Sarah Schmidt

En septiembre, un video adulterado del *Jornal Nacional*, el noticiero principal de la cadena Globo de televisión de Brasil, se propagó en las redes sociales. En el mismo, los presentadores William Bonner y Renata Vasconcellos mostraban los resultados de una encuesta de intención de voto para la Presidencia, pero los datos estaban invertidos al respecto de quién era el candidato favorito, tanto en los gráficos como en las intervenciones de los presentadores. Al día siguiente, el noticiero emitió una aclaración advirtiendo que el video se estaba utilizando para desinformar a la población y afirmó que se trataba de *deepfake*, o una ultrafalsificación, es decir, una técnica que se vale de la inteligencia artificial para realizar ediciones avanzadas de contenidos. Con la misma es posible, por ejemplo, alterar digitalmente el rostro de una persona o simular su voz, haciéndole hacer algo que no hizo o decir algo que no dijo.

En agosto, otro video del noticiero con una edición similar, que también invertía los resultados de una encuesta de intención de voto a la Presidencia, fue subido a la red social de videos TikTok, donde llegó a las 2.500.000 visualizaciones, según Projeto Comprova, una iniciativa que agrupa a periodistas de 43 medios de comunicación del país para chequear la desinformación.

"Puede ser que se haya utilizado alguna técnica de *deepfake* en esos videos, pero es necesario efectuar un análisis más minucioso. Para nosotros, lo que importa es saber que son falsos", dice el científico de la computación Anderson Rocha, director del Instituto de Computación de la Universidad de Campinas (Unicamp), en donde coordina el Laboratorio de Inteligencia Artificial (Recod.ai). El investigador ha venido estudiando formas de detectar adulteraciones maliciosas en fotos y videos, incluso en *deepfakes*, también llamadas medios sintéticos.

Otro caso ocurrió en marzo de este año, poco después de que comenzara la guerra entre Rusia y Ucrania, cuando el presidente ucraniano, Volodímir Zelenski fue víctima de *deepfake*. En las redes sociales circuló un video en el que él parecía pedirles a los ucranianos que depusieran las armas y regresaran a sus hogares, como si el país se estuviera rindiendo, lo que obligó a Facebook y YouTube a retirarlo de circulación tan pronto como se comprobó que era falso. En las imágenes, el rostro del presidente aparecía en un cuerpo que casi no se movía, vestido con una camiseta verde.

En algunos casos, como en los videos del *Jornal Nacional*, no es tan difícil darse cuenta de que habían sido alterados de alguna forma, porque las noticias originales se encuentran fácilmente disponibles para su verificación. Pero no siempre es así. Frente a los medios sintéticos, el aforismo "ver para creer" va perdiendo sentido y la propia inteligencia artificial podría ser una aliada.

"Por lo general, los videos sintéticos se elaboran en dos etapas: en primera instancia, mediante el uso de una plataforma de *deepfake*, para reemplazar rostros o realizar una sincronización de la boca, y después se procede a editarlos por medio de programas de edición de video", explica Rocha. Los que saben buscar suelen detectar alguna falla del programa utilizado para producir la falsificación: un juego de luces diferente o un contraste entre el video original y la cara nueva que se ha insertado.

Es como recortar un rostro de una foto y colocarlo sobre otra: el ángulo de iluminación y la forma en que la cámara captó las dos imágenes son diferentes. Estos rastros son pistas que se van dejando en el ca-

ILUSTRACIÓN FELIPE BRAZ CON FOTO DE ALEXA / PIXABAY

mino, que son identificadas mediante las técnicas de computación forense, un área de la investigación científica que ha crecido en los últimos años y de la cual Rocha forma parte.

Junto a colegas de la Universidad de Hong Kong, el investigador ha desarrollado un algoritmo que ayuda a detectar en forma simultánea en los videos si ha habido manipulaciones de rostros y, en caso de ser así, a determinar qué partes fueron alteradas. Puede haber sido un rostro completo, por ejemplo, o tan solo la boca, la zona de los ojos o el cabello. "El promedio de aciertos fue de un 88 % para los videos de baja resolución y de un 95 % para los que tenían una resolución superior", explica Rocha, sobre un conjunto total de 112.000 rostros sometidos a pruebas: la mitad verdaderos, la otra mitad manipulados a través de cuatro programas de *deepfake*. Este método también indica si la imagen fue generada desde cero, es decir, sin tener como base una fotografía real. Los resultados salieron publicados en abril de 2022 en la revista *Transactions on Information Forensics and Security*.

Según el científico de la computación, otros algoritmos desarrollados pueden detectar rasgos de alteraciones en los videos de *deepfake*, pero la mayoría trabajan con base en las pistas que han dejado los programas de manipulación más conocidos que, básicamente, se dividen en dos categorías: los que permiten el intercambio de rostros y aquellos que posibilitan la edición de las expresiones faciales. A un *software* de esta clase se lo conoce porque siempre deja alguna imperfección en la sincronización de la boca, y, por ello, se programa al algoritmo detector para que busque este error específico. "Esto conlleva un problema: si no conocemos el *software* de *deepfake*, resulta más difícil detectar esos rasgos. Y siempre están surgiendo aplicaciones nuevas", dice Rocha.

Por eso, él y sus colegas entrenaron al algoritmo que desarrollaron para que detecte las pistas sin suponer el conocimiento de la aplicación generadora de *deepfake*. "Trabajamos con la idea de que, independientemente del programa, este va a dejar un ruido, algo que no concuerda con el resto de la imagen". Este método trabaja en dos frentes: busca firmas de ruido, es decir, cambios sutiles en el borde de la cara, por ejemplo, y mapea la llamada firma semántica, que puede ser un defecto de color, textura o forma.

"El algoritmo automatiza el proceso que realiza un experto humano, consistente en buscar incoherencias, como los contrastes de luz", dice. "El paso siguiente consistirá en probarlo con videos falsos generados por una cantidad de programas mayor, para confirmar su potencial".

Este tipo de algoritmo detector puede utilizarse para diversos propósitos que involucren el uso malintencionado de *deepfakes*. Rocha forma parte de un proyecto internacional, llamado Semantic Forensics, junto con otros investigadores de las universidades de Siena y Politécnica de Milán, en Italia, y de Notre Dame, en Estados Unidos, que cuenta con el apoyo del Departamento de Defensa de Estados Unidos. El objetivo es desarrollar herramientas automatizadas que detecten estas manipulaciones. "Hemos visto casos de videos alterados de ejercicios militares de otros países, que multiplican la cantidad de misiles para mostrar un poderío bélico mayor que el real", relata.

Estos algoritmos también pueden ser de ayuda en los casos de *deepfakes* políticos, como en el episodio con el presidente ucraniano, o incluso pornográficos. La técnica se hizo famosa a finales de 2017, justamente utilizando películas de sexo. Por entonces, algunos usuarios de internet empezaron a insertar el rostro de celebridades del cine en escenas de películas con contenido sexual. Según una investigación realizada en septiembre de 2019 por DeepTrace Labs, una compañía neerlandesa de ciberseguridad, el 96 % de los videos de *deepfake* identificados en la red correspondía a pornografía no consentida. Las víctimas principales eran mujeres, sobre todo actrices, pero también se registraron casos de personas que no eran famosas. En julio de este año, la cantante Anitta también fue víctima de un *deepfake* porno. El video original ya había sido utilizado para producir otras

En un video falsificado, el presidente ucraniano, Volodímir Zelenski, les pedía a sus compatriotas que depusieran las armas

ES POSIBLE CAMBIAR EL ROSTRO DE UNA PERSONA O SIMULAR SU VOZ PARA LOGRAR QUE ESTA DIGA ALGO QUE NO DIJO

falsificaciones con el rostro de la actriz Angelina Jolie.

Según la periodista Cristina Tardáguila, directora de programas del Centro Internacional para Periodistas (ICFJ) y fundadora de Agência Lupa, que se especializa en la comprobación de sucesos, Brasil ya ha tenido que lidiar con *deepfakes* que tuvieron que desmentirse. Por ello, los programas que ayuden a detectar medios sintéticos pueden convertirse en aliados de los periodistas y verificadores de hechos, que trabajan contrarreloj. "Al lidiar con la desinformación, se necesita actuar con celeridad. Por eso, es importante invertir en inteligencia artificial, en herramientas que puedan ayudar a detectar y mapear con mayor rapidez este tipo de contenidos falsos. Así, podremos acortar el tiempo entre la propagación de contenido falso y la entrega de la verificación", analiza.

"Los ultrafalsos son la cumbre de las *fake news*. Tienen el potencial de poder engañar con mayor facilidad, porque, cuando se trata de un video, uno está viendo esa escena", dice la periodista Magaly Prado, quien cumple una pasantía posdoctoral en el Instituto de Estudios Avanzados de la Universidad de São Paulo (IEA-USP). "El audio, incluso, también puede generarse en forma sintética", dice la periodista, autora del libro *Fake news e inteligência artificial: O poder dos algoritmos na guerra da desinformação*, publicado en julio por Edições 70.

A su juicio, pese a ser menos recordados y habituales, los *deepfakes* exclusivamente en formato de audio tienen potencial para propagarse por plataformas como WhatsApp, una aplicación muy utilizada por los brasileños. Estos siguen una lógica similar a la de los videos: mediante aplicaciones accesibles que son cada vez mejores, puede simularse la voz de alguien. Las víctimas más fáciles son las personalidades públicas, cuya voz se encuentra ampliamente a disposición en internet. La técnica también puede utilizarse para perpetrar delitos económicos. "Ha habido casos como el de un empleado de una empresa de tecnología que recibió un correo de voz de un alto ejecutivo solicitándole que realizara una transferencia de fondos. Él sospechó, el caso fue analizado por una empresa de seguridad y se comprobó que se trataba de un mensaje elaborado por medio de inteligencia artificial", relata.

El periodista Bruno Sartori, director de la empresa FaceFactory, explica que la producción de *deepfakes* bien hechos, tanto de audio como de video, no resulta tan sencilla, por ahora. Esto es lo que él hace profesionalmente: su empresa se dedica a la creación de medios sintéticos para uso comercial y produce contenidos para programas humorísticos de los canales de televisión Globo y SBT.

En 2021, trabajó en un comercial para Samsung en el cual, la presentadora Maísa, ya adulta, interactuaba con ella misma cuando era una niña; esta última, creada a partir de técnicas de *deepfake*. La chiquilla virtual baila, juega y arroja una *notebook* hacia arriba. En otra ocasión, tuvo que insertar el rostro de un actor en un doble. "Para entrenar adecuadamente a la inteligencia artificial, es importante contar con un buen banco de imágenes y de audio de la persona que se pretende imitar. Los buenos programas que realizan procesamiento de alta calidad también deben disponer de configuraciones avanzadas. Caso contrario, pueden aparecer fallas perceptibles en el rostro, o bien, en el caso del audio, una voz robotizada", explica.

A su juicio, los videos manipulados del *Jornal Nacional* que tergiversan los datos de las encuestas no habrían llegado a ser alterados mediante el uso de inteligencia artificial. "Según mi análisis, en esos caso se hizo una edición común, que implica el corte y la inversión del orden de los audios. Se trata de lo que denominamos *shallowfake*. Empero, como están bien elaborados, su potencial para engañar a la gente es el mismo", analiza Sartori. Para él, en pocos años más, estos programas serán más ligeros, inteligentes y accesibles.

Hay algunas formas de protegerse contra la desinformación generada con ayuda de la tecnología. Una de ellas consiste en prestar atención a las licencias de uso y privacidad de las múltiples aplicaciones gratuitas que se utilizan de manera cotidiana, desde aquellas que solicitan acceso a las fotos del usuario para generar efectos divertidos, y pasando por las que pueden almacenar la voz. Según Rocha, de la Unicamp, muchas de ellas almacenan una gran cantidad de datos que podrían llegar a compartirse para otros propósitos, por ejemplo, entrenar programas de *deepfake*.

Otro punto importante es la educación mediática. "Por más que existan *software* que nos ayuden a detectar lo que es falso, el primer paso reside en desconfiar de aquello que se recibe en las redes sociales. Y comprobar las fuentes de información, investigarlas", concluye. ■

Proyecto

Déjà vu: coherencia temporal, espacial y de caracterización de datos heterogéneos para el análisis y la interpretación de la integridad (nº 17/12646-3); **Modalidad** Proyecto Temático; **Investigador responsable** Anderson Rocha; **Inversión** R$ 1.912.168,25.

Artículo científico

KONG, C. *et al.* Detect and locate: Exposing face manipulation by semantic- and noise-level telltales. **Transactions on Information Forensics and Security**. v. 17. abr. 2022.

Libro

PRADO, M. **Fake news e inteligência artificial: O poder dos algoritmos na guerra da desinformação**. São Paulo: Edições 70, 2022.

REPRODUCCIÓN