Este documento ha sido traducido por la Biblioteca del Congreso de la República con fines meramente informativos para los usuarios de la institución. Se trata de una traducción no oficial del texto en inglés elaborado por Kelley M. Sayler, Analista de Tecnología Avanzada y Seguridad Global, y Laurie A. Harris, Analista de Ciencia y Tecnología Política del Congressional Research Service (CRS).

Título del documento:

Inglés: Deep Fakes and National Security
N° de páginas: 3
Enlace**:** https://crsreports.congress.gov/product/pdf/IF/IF11333#:~:text=Deep%20fakes%20could%
Fecha de documento: 17 de abril del 2023

Español: Falsificaciones profundas y seguridad nacional
N° de páginas: 4
Fecha de documento: setiembre del 2023*

Institución: El Servicio de Investigación del Congreso (CRS) de Estados Unidos brinda análisis políticos y jurídicos a las comisiones y a los miembros de la Cámara de Representantes y del Senado, independientemente de su afiliación partidista. Como órgano del poder legislativo dentro de la Biblioteca del Congreso, el CRS es un recurso valioso y respetado en el Capitolio desde hace más de un siglo. El CRS es conocido por sus análisis autorizados, confidenciales, objetivos y no partidistas. Su máxima prioridad es garantizar que el Congreso tenga acceso las 24 horas del día a las mejores ideas de la nación.

* *N.T.: Esta traducción ha sido realizada por la Biblioteca del Congreso (traductora: MPZ).*

Falsificaciones profundas y seguridad nacional

Las «falsificaciones profundas» (*deep fakes*) —término que apareció por primera vez en el 2017 para describir falsificaciones realistas de fotos, audios, vídeos y de otro tipo, generadas con tecnologías de inteligencia artificial (IA)— podrían plantear diversos problemas de seguridad nacional en los próximos años. A medida que estas tecnologías sigan madurando, podrían tener implicaciones significativas para la función de control del Congreso, para las autorizaciones y asignaciones a la defensa de Estados Unidos y para la regulación de las plataformas de medios sociales.

¿Cómo se crean las falsificaciones profundas?

Aunque las definiciones varían, en general, las falsificaciones profundas se suelen describir como falsificaciones creadas utilizando técnicas de aprendizaje automático (*machine learning*, ML) —un subcampo de la IA—, especialmente redes generativas antagónicas (GAN). En el proceso de GAN, dos sistemas de ML, llamados redes neuronales, se entrenan compitiendo entre sí. La primera red, o la generadora, se encarga de crear datos falsificados —como fotos, grabaciones de audio o metrajes de video— que reproducen las propiedades del conjunto de datos original. La segunda red, o la discriminadora, se encarga de identificar los datos falsificados. En función de los resultados de cada iteración, la red generadora se ajusta para crear datos cada vez más realistas. Las redes siguen compitiendo —a menudo durante miles o millones de iteraciones— hasta que la generadora mejora su rendimiento al punto de que la discriminadora ya no puede distinguir entre datos reales y falsos.

Aunque la manipulación de los medios de comunicación no es un fenómeno nuevo, el uso de la IA para generar falsificaciones profundas es motivo de preocupación, porque los resultados son cada vez más realistas, se crean con rapidez y son baratos gracias a la posibilidad de alquilar potencia de procesamiento a través de la computación en la nube. Así, incluso los operadores no cualificados podrían descargar las herramientas de *software* necesarias y, utilizando datos disponibles públicamente, crear contenidos falsificados cada vez más convincentes.

¿Cómo podrían utilizarse las falsificaciones profundas?

La tecnología de falsificación profunda se ha popularizado con fines de entretenimiento: por ejemplo, los usuarios de las redes sociales que insertan al actor Nicholas Cage en películas en las que no apareció originalmente y un museo que genera una exposición interactiva con el artista Salvador Dalí. Las tecnologías de falsificación profunda también se han utilizado con fines benéficos. Por ejemplo, investigadores médicos han informado del uso de GAN para sintetizar imágenes médicas falsas con el fin de entrenar algoritmos de detección de enfermedades raras y minimizar los problemas de privacidad de los pacientes.

Sin embargo, las falsificaciones profundas podrían utilizarse con fines nefastos. Adversarios del Estado o individuos con motivaciones políticas podrían difundir vídeos falsificados de funcionarios electos u otras figuras públicas haciendo comentarios incendiarios o comportándose de forma inapropiada. Esto podría, a su vez, erosionar la confianza pública, afectar negativamente al discurso público o incluso influir en unas elecciones.

De hecho, la Comunidad de Inteligencia de Estados Unidos concluyó que Rusia participó en amplias operaciones de influencia durante las elecciones presidenciales del 2016 para «socavar la fe pública en la proceso democrático estadounidense, denigrar a la Secretaria de Estado Hillary Clinton y perjudicar sus posibilidades de elección para la presidencia». Asimismo, en marzo del 2022, el presidente ucraniano Volodymyr Zelensky

anunció que un vídeo publicado en las redes sociales —en el que parecía ordenar a los soldados ucranianos que se rindieran a las fuerzas rusas— era una falsificación profunda. Aunque los expertos señalaron que esta falsificación profunda no era especialmente sofisticada, en el futuro, las falsificaciones convincentes de audio o vídeo podrían reforzar las operaciones de influencia maliciosas.

Las falsificaciones profundas también podrían utilizarse para avergonzar o chantajear a funcionarios electos o personas con acceso a información clasificada. Ya existen pruebas de que agentes de inteligencia extranjeros han utilizado fotos falsas para crear cuentas falsas en redes sociales desde las que han intentado reclutar fuentes. Algunos analistas han sugerido que las falsificaciones profundas también podrían utilizarse para generar contenidos incendiarios —como vídeos convincentes de militares estadounidenses implicados en crímenes de guerra— destinados a radicalizar a la población, reclutar terroristas o incitar a la violencia. El artículo 589F de la Ley de Autorización de la Defensa Nacional para el año fiscal 2021 (P.L. 116-283) ordena al Secretario de Defensa que lleve a cabo una evaluación de inteligencia de la amenaza que suponen las falsificaciones profundas para los miembros de las fuerzas armadas y sus familias, incluyendo una evaluación de la madurez de la tecnología y cómo podría utilizarse para llevar a cabo operaciones de información.

Además, las falsificaciones profundas podrían producir un efecto que los profesores Danielle Keats Citron y Robert Chesney han denominado el «beneficio del mentiroso»; se trata de la noción de que los individuos podrían negar con éxito la autenticidad de un contenido genuino —especialmente si muestra un comportamiento inapropiado o delictivo— alegando que el contenido es una falsificación profunda. Citron y Chesney sugieren que el «beneficio del mentiroso» podría hacerse más poderoso a medida que prolifere la tecnología de falsificación profunda y aumente su conocimiento por el público.

Algunos informes indican que ya se han utilizado estas tácticas con fines políticos. Por ejemplo, los adversarios políticos del Presidente de Gabón, Ali Bongo, afirmaron que un vídeo que pretendía demostrar su buena salud y sucompetencia mental era una falsificación profunda, y más tarde lo citaron como parte de la justificación de un intento de golpe de Estado. Los expertos externos no pudieron determinar la autenticidad del vídeo, pero uno de ellos señaló que «en cierto modo, no importa si [el vídeo es] falso... Puede utilizarse simplemente para socavar la credibilidad y sembrar la duda».

¿Cómo detectar falsificaciones profundas?

Hoy en día, las falsificaciones profundas pueden detectarse a menudo sin herramientas de detección especializadas. Sin embargo, la sofisticación de la tecnología está progresando rápidamente hasta un punto en el que la detección humana sin ayuda será muy difícil o imposible. Mientras que la industria comercial ha estado invirtiendo en herramientas automatizadas de detección de falsificaciones profundas, esta sección describe las inversiones y actividades del Gobierno de Estados Unidos.

La Ley de Identificación de los Productos de las Redes Generativas Antagónicas (P.L. 116-258) encarga a la National Science Foundation (NSF) [Fundación Nacional de Ciencias] y al National Institute of Standards and Technology (NIST) Instituto Nacional de Normas y Tecnologías] a que apoyen la investigación sobre las GAN. En concreto, la NSF debe apoyar la investigación sobre contenidos manipulados o sintetizados y la autenticidad de la información, y el NIST debe apoyar la investigación para el desarrollo de medidas y normas necesarias para desarrollar herramientas que permitan examinar la función y los resultados de las GAN u otras tecnologías que sintetizan o manipulan contenidos.

Además, DARPA (la Agencia de Proyectos de Investigación Avanzados de Defensa) ha tenido dos programas dedicados a la detección de falsificaciones profundas: Media Forensics (MediFor) y Semantic Forensics (SemaFor). MediFor, que concluyó en el año fiscal 2021, debía desarrollar algoritmos para evaluar automáticamente la integridad de fotos y vídeos, y proporcionar a los analistas información sobre cómo se generaron los contenidos falsificados. Al parecer, el programa exploró técnicas para identificar las incoherencias audiovisuales presentes en las falsificaciones profundas, incluidas las incoherencias en los píxeles (integridad digital), las incoherencias con las leyes de la física (integridad física) y las incoherencias con otras fuentes de información (integridad semántica). Se espera que las tecnologías MediFor pasen a los mandos operativos y a la comunidad de inteligencia.

SemaFor pretende basarse en las tecnologías de MediFor y desarrollar algoritmos que detecten, atribuyan y caractericen automáticamente (es decir, identifiquen como benignos o maliciosos) diversos tipos de falsificaciones profundas. El objetivo de este programa es catalogar las incoherencias semánticas —como los pendientes no coincidentes que se ven en la imagen generada por GAN de la **figura 1**, o los rasgos faciales o fondos inusuales— y priorizar las presuntas falsificaciones profundas para su revisión humana. DARPA solicitó 28,9 millones de dólares para SemaFor en el año fiscal 2023, 7,9 millones más que en el año fiscal 2022. Las tecnologías desarrolladas tanto por SemaFor como por MediFor están destinadas a mejorar las defensas contra las operaciones de información antagónica.

Figura 1. Ejemplo de incoherencia semántica en una imagen generada por una GAN

Fuente: https://www.darpa.mil/news-events/2019-09-03a

Consideraciones políticas

Algunos analistas han señalado que las herramientas de detección basadas en algoritmos podrían conducir al juego del gato y el ratón, en el que los generadores de falsificaciones profundas se actualizan rápidamente para subsanar los fallos identificados por las herramientas de detección. Por este motivo, sostienen que las plataformas de redes sociales, además de desplegar herramientas de detección de falsificaciones profundas, pueden necesitar ampliar los medios de etiquetado o autenticación de contenidos. Esto podría incluir que los usuarios identifiquen la hora y el lugar en que se que se originó el contenido o que etiqueten el contenido editado como tal.

Otros analistas han expresado su preocupación por que la regulación de la tecnología de falsificación profunda pueda imponer una carga indebida a las plataformas de medios sociales o dar lugar a restricciones inconstitucionales de la libertad de expresión, y la expresión artística. Estos analistas han sugerido que la legislación vigente es suficiente para gestionar el uso malintencionado de las falsificaciones profundas. Algunos expertos han afirmado que responder únicamente con herramientas técnicas será insuficiente y que, en su

lugar, habría que centrarse en la necesidad de educar al público sobre las falsificaciones profundas y minimizar los incentivos para los creadores de falsificaciones profundas maliciosas.

Posibles preguntas para el Congreso

- ¿El Departamento de Defensa, el Departamento de Estado y la comunidad de inteligencia disponen de información adecuada sobre el estado de la tecnología extranjera de falsificación profunda y las formas en que esta tecnología puede utilizarse para dañar la seguridad nacional de Estados Unidos?
- ¿Hasta qué punto están maduros los esfuerzos de DARPA por desarrollar herramientas automatizadas de detección de falsificaciones profundas? ¿Cuáles son las limitaciones del enfoque de DARPA, y si se requieren esfuerzos adicionales para garantizar que las falsificaciones profundas malintencionadas no perjudiquen la seguridad nacional de Estados Unidos?
- ¿Son adecuadas las inversiones federales y los esfuerzos de coordinación, a través de las agencias de defensa y de no defensa, y con el sector privado, para abordar las necesidades de investigación y desarrollo, y las preocupaciones de seguridad nacional con respecto a las tecnologías de falsificación profunda?
- ¿Cómo deben equilibrarse las consideraciones de seguridad nacional con respecto a las falsificaciones profundas con la protección de la libertad de expresión, la expresión artística y los usos beneficiosos de la tecnología subyacente?
- ¿Debería exigirse a las plataformas de las redes sociales que autentifiquen o etiqueten los contenidos? ¿Debería exigirse a los usuarios que faciliten información sobre la procedencia de los contenidos? ¿Qué efectos secundarios podría tener esto para las plataformas de medios sociales y la seguridad y privacidad de los usuarios?
- ¿Hasta qué punto y de qué manera, si acaso, deben las plataformas de redes sociales y los usuarios son responsables de la difusión y las repercusiones de los contenidos maliciosos de falsificaciones profundas maliciosas?
- ¿Qué esfuerzos, en su caso, debería emprender el Gobierno de Estados Unidos para garantizar que el público esté informado sobre las falsificaciones profundas?

Productos CRS

Informe CRS R45178, *Artificial Intelligence: Background, Selected Issues, and Policy Considerations*, por Kelley M. Sayler

Informe CRS R46795, *Artificial Intelligence and National Security*, por Laurie A. Harris

Informe CRS R45142, *Information Warfare: Issues for Congress*, por Catherine A. Theohary

Kelley M. Sayler, Analista de Tecnología Avanzada y Seguridad Global

Laurie A. Harris, Analista de Política Científica y Tecnológica

Descargo de responsabilidad