Resistances
Journal of the Philosophy of History

Los riesgos de los deepfakes para la democracia y las perspectivas de regulación

The risks of deepfakes for democracy and the prospects for regulation

Efraín Fandiño López

Resumen

Este artículo realiza un análisis del impacto de los deepfakes en el marco de los procesos democráticos, enfocándose en cómo estas sofisticadas manipulaciones digitales pueden erosionar la confianza pública, interferir en elecciones y generar polarización. Se estudia la progresión de las ultra falsificaciones desde su creación hasta alcanzar un alto grado de sofisticación, destacando cómo los avances en tecnología de inteligencia artificial han potenciado su realismo y complicado su detección. Se hará alusión a varios casos en Francia y Colombia, con el ánimo de examinar las repercusiones adversas de los deepfakes en la representación política y la percepción por parte del público en general. El documento propone la adopción de marcos regulatorios y medidas que trasciendan las medidas punitivas tradicionales, abogando por estrategias preventivas y la promoción de la transparencia en el empleo de tecnologías de inteligencia artificial. Esta aproximación puede ayudar a promover los valores democráticos y un uso responsable de la IA.

Palabras clave: Democracia; Deepfakes; Desinformación; Regulación de la IA; Inteligencia artificial; Transparencia en IA; Ética; Marcos regulatorios.

Efraín Fandiño López
Investigador Independiente | Bogotá | Colombia | Efandino91@outlook.com
https://orcid.org/0000-0003-0169-7850

http://doi.org/10.46652/resistances.v5i10.167
ISSN 2737-6230
Vol. 5 No. 10 julio-diciembre 2024, e240167
Quito, Ecuador

Enviado: agosto, 15, 2024
Aceptado: octubre, 16, 2024
Publicado: noviembre, 11, 2024
Publicación Continua

RELIGACIÓN CICSHAL
Centro de Investigaciones en Ciencias Sociales y Humanidades desde América Latina

Abstract

This article examines the impact of deepfakes in a democratic context, focusing on how these digital manipulations can undermine public trust, manipulate political information, interfere with electoral processes, and exacerbate polarization. It analyzes the evolution of deepfakes from their inception to their current sophistication, illustrating how artificial intelligence technology has enhanced their realism and made them more difficult to detect. The paper examines the harmful effects of deepfakes on political representation, polarization, and public perception. It draws on case studies in France and Colombia. The paper argues for regulatory strategies that go beyond regulatory measures and promote preventive approaches and transparency in the use of artificial intelligence technologies. It underlines the crucial role of cooperation with platforms and the implementation of educational and technological policies that promote a critical understanding of digital media among citizens. This comprehensive approach aims to protect democratic values and ensure the responsible use of artificial intelligence.

Keywords: Democracy; Deepfakes; Desinformation; AI Regulation; Artificial intelligence; Transparency in AI; Ethics, Regulatory Frameworks.

Introducción

La crisis de la democracia es un leitmotiv recurrente en nuestras sociedades. Desde las contribuciones de Crozier et al. (1975), hasta las reflexiones de Guillermo O'Donnell (2007), diversos académicos han destacado las tensiones y problemas inherentes al sistema político democrático. Estas resistencias incluyen la pérdida de la legitimidad de los liderazgos, la carencia de representatividad, la falta de objetivos comunes que conlleva el choque de intereses particulares antagónicos, y las persistentes desigualdades sociales. No obstante, el malestar social, si bien es una expresión legítima de las necesidades y demandas de los ciudadanos, también ha sido cooptado por aquellos que buscan manipularlo para beneficio propio, a través de estrategias de propaganda y desinformación que creen un clima de inestabilidad y deslegitimación del poder establecido, con el objetivo final de acceder a él.

En este contexto, las democracias se han visto confrontadas en los últimos años con desafíos significativos emergentes resultado de la digitalización. Como lo señalan varios autores, este desarrollo ha transformado radicalmente los modos de diseminación y consumo de noticias, facilitando la expansión de la desinformación a través de noticias falsas que erosionan la confianza en las instituciones democráticas (McKay y Tenove, 2020). Asimismo, otros registran la manipulación de algoritmos en intentos recientes por dirigir la comunicación política a escala global (Woolley y Howard, 2018), generando como consecuencia burbujas de filtro que limitan la exposición a perspectivas diversas e intensifican la polarización en las redes sociales (Chitra y Musco, 2020).

Aunque los casos mencionados siguen planteando desafíos para los ciudadanos y los gobiernos, en años recientes, la desinformación ha entrado en una nueva era con la aparición de los *deepfakes*, o ultrafalsificaciones, creados mediante el uso de inteligencia artificial (IA). Esta innovación

tecnológica ha suscitado una creciente preocupación entre los gobiernos democráticos globales, dada su capacidad para elevar la mentira a niveles sin precedentes.

En este contexto, sobresale la siguiente problemática: ¿De qué manera pueden los *deep fakes* contribuir a la erosión de la confianza en la democracia, y cuáles son las estrategias regulatorias efectivas para atenuar dichos efectos?

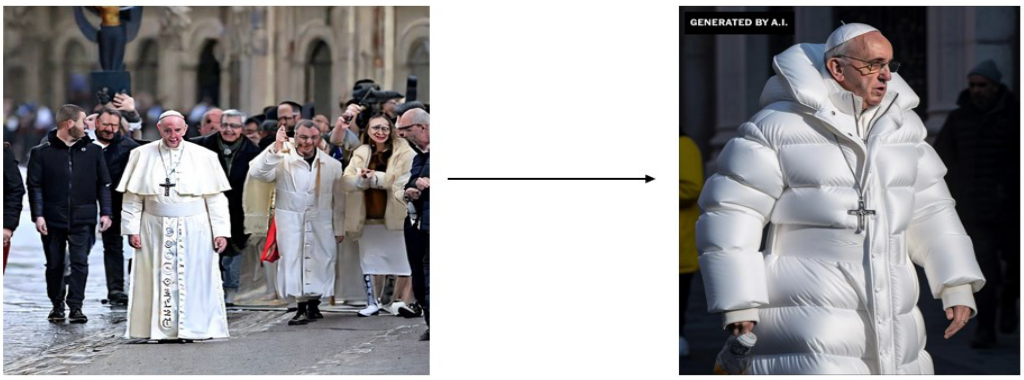
Con el ánimo de proporcionar una respuesta a esta problemática, este documento expone en el capítulo 2 la evolución de las ultrafalsificaciones. El Capítulo 3 investiga el impacto de los *deepfakes* en las democracias contemporáneas, analizándolos como agentes de desinformación. Posteriormente, los capítulos 4 y 5 examinan la imperiosa necesidad de implementar un marco regulatorio que mitigue los riesgos asociados con las ultrafalsificaciones, tanto desde una perspectiva general como específicamente en el contexto democrático. Este análisis incluye una exploración detallada de las soluciones regulatorias pertinentes, enfocándose en su aplicación tanto en el ámbito electoral como en la protección de las instituciones democráticas.

La evolución de los Deepfakes

Según Meredith Somers (2020), la ultrafalsificación emergió en el horizonte digital entre 2017 y 2018, constituyendo una evolución en la trayectoria de las herramientas digitales de desinformación. En esencia, un *deepfake* se define como una creación hecha mediante inteligencia artificial, que se caracteriza por ser un conjunto de datos sintéticos, creados mediante la modificación o manipulación automática de datos preexistentes (imágenes, textos, sonidos o videos), con el propósito explícito de engañar al observador. Esta manipulación digital, según el diccionario, se extiende desde la modificación de imágenes y vídeos, tales como la sustitución de rostros en fotografías y la simulación de voces en grabaciones, hasta la generación de imágenes, textos o audios que replican con exactitud fotografías o material de naturaleza audiovisual (Merriam-Webster).

Desde su aparición en 2017, la tecnología detrás de los *deepfake* ha experimentado avances vertiginosos. Esta evolución se debe a las mejoras técnicas de inteligencia artificial (como las redes neuronales convolucionales) las cuales han transformado estas falsificaciones, que inicialmente presentaban errores visuales y de sincronización, en imitaciones tan precisas que desafían la detección por observadores no especializados. Para ilustrar la evolución de la ultrafalsificación, es instructivo examinar comparativamente una imagen que se hizo viral del Papa Francisco en 2023, con otra generada automáticamente en 2022. El análisis detallado de ambas imágenes revela diferencias significativas en términos de realismo y calidad de la simulación.

Figura 1. Comparativa de imágenes del Papa Francisco generadas por IA



Fuente: Stable Diffusion (2022); The New York Times (2023).

Nota. La primera, creada con Stable Diffusion en 2022, y la segunda, documentada por The New York Times en 2023.

La imagen del año 2022 presenta múltiples anomalías, como irregularidades en la representación del rostro, las manos y el vestuario del Papa, acompañadas de expresiones faciales incongruentes y texturas de piel antinaturales, denotando su naturaleza artificial. En contraste, la fotografía del año 2023, reseñada por Huang en el New York Times, muestra un notable avance en el realismo, con un retrato del Papa Francisco fácilmente confundido con una imagen auténtica, evidenciado por un rostro y unas manos similares a los de la persona real. Estas comparaciones subrayan la rápida progresión y refinamiento de los *deepfakes* en los últimos años, el cual hace más difícil la diferenciación entre contenido genuino y fabricado.

En principio, las ultra falsificaciones no son intrínsecamente malas. Estas abrieron un mundo de posibilidades en el ámbito creativo (Silbey y Hartzog 2019). En el cine, por ejemplo, la tecnología ha permitido revivir personajes icónicos como la Princesa Leia en "Rogue One: Una historia de Star Wars", rejuveneciendo a la actriz Carrie Fisher de forma impecable. En el marketing, marcas como Zalando o Cadbury han utilizado ultra falsificaciones para crear campañas publicitarias interesantes (Mediatropy). No obstante, la percepción negativa de esta clase de creaciones sintéticas proviene de los usos ilícitos y maliciosos que se le han conocido. Es de recordar que el público no especializado conoció los *deepfakes* por los montajes pornográficos (Harris, 2019). En efecto, individuos, de forma anónima, trucaban videos pornográficos de tal suerte que superponían rostros de actrices reconocidas sin su consentimiento, creando una apariencia real. Esta práctica se extendió a mujeres fuera del ámbito público, convirtiendo a los *deepfakes* en un instrumento para humillar y deshonrar a mujeres, independientemente de su estatus público. Como se analizará en secciones posteriores, estas prácticas se han extendido al ámbito político y democrático, empleando las ultra falsificaciones como herramienta para utilizar la mentira con el ánimo de silenciar a figuras

públicas, influir en procesos electorales, promover la polarización y deformar la percepción del ciudadano sobre algún tema de interés colectivo.

Es así como los riesgos asociados al uso de estas falsificaciones generadas mediante inteligencia artificial han llevado a numerosos expertos y ciudadanos a considerar imprescindible la implementación de un marco legal que establezca límites claros y sanciones específicas para el mal uso de estas tecnologías avanzadas. Sin embargo, antes de entrar en el tema de la regulación, es menester explorar las repercusiones de este tipo de falsificaciones en el discurso democrático y la política.

La ultra falsificación y sus repercusiones en las democracias modernas

El engaño en el contexto político y democrático ha existido desde tiempos antiguos, de tal suerte que la desconfianza hacia políticos no es un fenómeno nuevo, sino un eco persistente a través de los siglos, una queja que resuena en el tiempo, tan antigua como el arte de gobernar. Tal como lo destaca Bellamy (2020), desde la época de Platón hasta la era contemporánea marcada por la influencia de autores como Arendt, este fenómeno ha adoptado diversas formas de manifestación, que abarcan desde la ocultación de actos individuales hasta la manipulación de la información en ámbitos tanto públicos como privados.

No obstante, las llamadas ultra falsificaciones representan una nueva amenaza para las democracias, fundamentalmente por cómo erosionan la confianza en el tejido social y político, puesto que al crear contenidos falsos y virales pero verosímiles, se manipula la percepción pública, sembrando dudas y desinformación. Máxime, teniendo en cuenta que su facilidad de uso permite a personas sin experiencia técnica crear videos manipulados convincentes que se difunden rápidamente a través de las redes sociales. Adicional a lo anterior, autores como Farid (2021), han puesto de presente que, un factor que aumenta el riesgo en las democracias es el hecho de que, al ser un fenómeno tan reciente, no existen herramientas efectivas de detección de este tipo de contenidos que sean efectivas para luchar contra la desinformación.

Bajo esta perspectiva, un informe del Foro Económico Mundial (2024), considera que la desinformación es uno de los riesgos globales más severos y de rápido crecimiento. El texto señala que tanto la desinformación, como la información errónea, potenciadas por la inteligencia artificial, están aumentando rápidamente, lo que agrava la polarización a escala social y política. Esta tendencia es particularmente preocupante en el contexto de elecciones importantes en varios países, donde la propagación de información falsa puede socavar la legitimidad de los gobiernos elegidos e incluso, podría llegar al punto de desencadenar disturbios civiles.

En el marco del espacio político, existen dos tipos de uso de la ultrafalsificación: el justo y el perjudicial. Un uso legítimo de los *deepfakes* se manifestó en una parodia creada por un individuo

denominado *The French Faker* que involucró al presidente de Francia, Emmanuel Macron, en la cual se le representaba falsamente renunciando a su cargo mediante una alocución apócrifa. El citado producto audiovisual demuestra que las ultrafalsificaciones pueden fungir como una forma de sátira o parodia, contribuyendo así a la crítica política y social dentro de un marco ético, siempre y cuando se deje de presente que se trata de un producto sintético creado mediante inteligencia artificial.

Figura 2. Captura de pantalla de la falsa alocución presidencial del presidente Macron, creada mediante IA por un individuo que se denomina "French Maker".

Fuente: French Maker (s. f.)

Respecto de los usos maliciosos, podemos citar dos casos recientes que ocurrieron en Colombia y Francia que representan las consecuencias negativas del mal uso de los *deepfakes*. El primero, ocurrió durante las elecciones regionales de 2023 del país suramericano. Aunque este proceso electoral no presentó mayores novedades en términos de resultados, se destacó por un hecho registrado por el sitio Colombiacheck (2023), especializado en la verificación de desinformación en la web: dos candidatos populares, quienes resultaron electos por los ciudadanos en grandes urbes del país, fueron víctimas de manipulación digital. Mediante diferentes técnicas relacionadas con inteligencia artificial, se crearon montajes auditivos que imitaban con alta fidelidad sus voces, atribuyéndoles falsamente declaraciones que nunca pronunciaron. El propósito era, de manera evidente, menoscabar la credibilidad de los candidatos ante sus electores mediante la creación de mensajes falsificados en los que los distintos políticos admitían haber participado en sobornos inexistentes.

En la misma línea, se identificaron dos ejemplos de *deepfakes* utilizados en Francia con propósitos específicos: socavar la imagen del presidente Macron y difundir desinformación relacionada con el conflicto en Ucrania. Así, por un lado, se generaron videos trucados en los que aparecía un Macron juvenil participando en danzas dentro de un establecimiento de ambiente homosexual (Reuters). Adicionalmente, se manipuló material audiovisual del medio France 24 para fabricar la noticia de la cancelación de un viaje presidencial a Kiev, atribuyéndola a un inexistente complot ucraniano con intenciones de atentar contra la vida del mandatario francés (AFP). De este modo, se ejemplifica como individuos con intenciones maliciosas recurrieron a falsificaciones elaboradas

mediante inteligencia artificial con los propósitos de deteriorar la imagen personal del presidente ante los votantes conservadores y, en el segundo caso, influir en la percepción pública acerca de un asunto tan delicado como la guerra en Ucrania.

Es así como la desinformación a través de ultra falsificaciones, conduce a la emergencia del fenómeno denominado por Pawelec (2022), como "*deterioro de la confianza informativa*". Según esta autora, la capacidad de fabricar declaraciones falsas atribuidas a figuras públicas, modificar eventos o situaciones reales y generar escenarios ficticios mediante el uso de imágenes, videos o textos apócrifos, podría resultar en el debilitamiento de los fundamentos de la verdad y la objetividad. Estos pilares son esenciales para el funcionamiento efectivo de las democracias. Además, es preciso añadir que tales prácticas pueden provocar una erosión de la confianza ciudadana hacia las instituciones y los individuos que las lideran. Esto ocurre tanto por una desconfianza generalizada hacia cualquier información relacionada con figuras públicas o las entidades que representan, como por la credibilidad otorgada a información engañosa con intenciones maliciosas.

Desde esta óptica, nos enfrentamos a dos cuestiones de notable relevancia jurídica: en primer lugar, la vulneración de los derechos de los individuos, como pueden ser el derecho a la imagen y al buen nombre. Esta violación se materializa cuando se utiliza el rostro o la voz de una persona, en este caso representando una institución del Estado o como candidato de elección popular, para atribuirle declaraciones falsas que comprometen su reputación y credibilidad. En segundo término, se observa una transgresión colectiva de los derechos de los ciudadanos, ejemplificada por la vulneración al derecho a ser informado, tal como se reconoce en sistemas jurídicos como el colombiano. Esta prerrogativa comprende desde la posibilidad de acceder a información veraz hasta la disponibilidad de información pública de calidad, condiciones importantes para una participación del ciudadano en el ámbito democrático (Pulido et al., 2013). Tales violaciones no solo afectan a los individuos de manera aislada, sino que comprometen la integridad del tejido social y el funcionamiento democrático en su conjunto. En este contexto, no es extraño que se haya puesto sobre la mesa la necesidad de regulación de los *deepfakes* como veremos en el acápite siguiente.

La regulación de los deepfakes

Previo a abordar la cuestión de la regulación de las ultra falsificaciones, resulta imperativo realizar algunas consideraciones respecto a la reglamentación de las cada vez menos nuevas tecnologías y sus aplicaciones. El derecho, en su incesante fluir, se halla en un proceso constante de evolución, moldeándose a medida que emergen nuevos fenómenos sociales. Como resultado de esta adaptabilidad intrínseca, los operadores jurídicos adoptan constantemente medidas destinadas a modificar comportamientos sociales, por ejemplo, mediante la implementación de nuevas sanciones, o bien, a fomentar actividades específicas, como lo evidencia la instauración de exenciones fiscales para ciertas actividades (Commaille 2015 y Ost 2016).

Bajo este tenor, es un hecho de que los significativos avances tecnológicos desarrollados durante los últimos setenta años han dado lugar a la creación y uso de nuevas tecnologías que han transformado radicalmente los modos de vida de innumerables ciudadanos a nivel global. Así, la invención y masificación de la computadora personal en los años 80, la proliferación de internet en los 2000, la creación de plataformas digitales en la segunda década de los 2000 y, más recientemente, el desarrollo de la inteligencia artificial, han pasado a formar parte del objeto de estudio del derecho. Esto se debe a su notable influencia y sus consecuencias en las relaciones interpersonales y transnacionales entre individuos de distintos estados. En consecuencia, los operadores jurídicos han emprendido esfuerzos para delimitar el uso de estas tecnologías, principalmente mediante la implementación de normas de carácter jurídico, que abordan exhaustivamente tanto los impactos como las consecuencias legales que surgen del uso y desarrollo tecnológico. Este enfoque regulatorio busca, en teoría, establecer un marco legal que asegure la adaptación ética y responsable de las innovaciones tecnológicas, de tal suerte que se protejan los derechos de los individuos, fomentando al mismo tiempo un desarrollo tecnológico sostenible y equitativo.

Ahora bien, la idea de regular las nuevas tecnologías, atractiva en teoría, enfrenta desafíos significativos en la práctica, principalmente en determinar el momento óptimo para su regulación. Esta cuestión no es trivial, dado que una regulación estricta puede obstaculizar el desarrollo tecnológico, mientras que su ausencia puede facilitar abusos (Reins, 2019). Al respecto, la profesora Crootof (2019), propone varios criterios para evaluar la necesidad de regular una tecnología. Primero, Crootof sugiere que debe haber una "disrupción" causada por la tecnología, que puede manifestarse de diversas maneras: una modificación sustancial en la modalidad de aplicación del derecho, la presencia de ambigüedades jurídicas, la aparición de situaciones de incertidumbre en la aplicación o extensión de las normas vigentes, o el menoscabo de los supuestos o principios fundamentales que sustentan un régimen jurídico. Una vez identificada la naturaleza de la disrupción, indica la profesora que se debe determinar si nuevas interpretaciones de las leyes existentes pueden abordarla, si se requieren revisiones explícitas a las leyes vigentes, o si es necesario crear nuevas leyes ante la insuficiencia de los marcos legales existentes. Posteriormente, es crucial, según la autora establecer la urgencia o el 'timing' de la regulación. Para ello, se deben considerar varios aspectos: el potencial de daño irreversible si la tecnología no se regula con prontitud; los beneficios de un enfoque de espera y observación, que puede ser apropiado cuando los riesgos e impactos de la tecnología no se comprenden completamente; la aplicación del principio de precaución, especialmente cuando la tecnología plantea riesgos serios que podrían causar daños irreversibles; y la viabilidad de regulaciones proactivas que puedan prevenir daños significativos mientras se permiten realizar los beneficios de la tecnología.

Estas consideraciones son, para empezar, aplicables al ámbito de la inteligencia artificial, la tecnología que facilita la creación de ultra falsificaciones. La IA puede causar disrupciones en múltiples frentes. Por ejemplo, podría introducir ambigüedades legales (por ejemplo, en los derechos

de autor de las obras generadas por IA) o generar cambios en normas existentes (como en la privacidad y el manejo de datos personales). También puede cuestionar los principios fundamentales de la responsabilidad, al ser difícil determinar quién es culpable cuando un sistema autónomo toma decisiones erróneas (Zech, 2021). En muchos casos, la legislación actual no abarca específicamente los desafíos presentados por la IA, lo que puede requerir nuevas interpretaciones legales o incluso la creación de nuevas leyes para abordar efectivamente las cuestiones de responsabilidad, seguridad y ética. La implementación de regulaciones proactivas podría ayudar a prevenir daños significativos relacionados con la IA mientras se exploran y se realizan sus beneficios potenciales. Esto podría incluir, por ejemplo, reglamentos sobre transparencia en algoritmos y procesos de toma de decisiones, requisitos de seguridad para sistemas autónomos, y directrices éticas para el desarrollo y despliegue de IA según Vokinger y Gasser (2021).

En este contexto, es relevante señalar que, desde diversas partes del mundo, varios estados han emprendido esfuerzos para regular esta tecnología. En la Unión Europea, fue aprobada hace unos meses una reglamentación europea de la inteligencia artificial basada en riesgos. En el momento en que se escriben estas palabras, existen en el parlamento colombiano 7 proyectos de ley de regulación de esta tecnología. En Estados Unidos, podemos rescatar la Orden Ejecutiva número 14110 decretada por el presidente Biden referente al Desarrollo y Uso Seguros y Fiables de la Inteligencia Artificial, en cuya reglamentación existen normas que buscan garantizar la seguridad en la creación y uso de sistemas de inteligencia artificial. Diversas de estas normativas han sido acogidas favorablemente por aquellos que reconocen la necesidad de establecer restricciones ante los usos abusivos de la inteligencia artificial. No obstante, también han sido objeto de críticas por parte de quienes argumentan que es prematuro implementar una regulación sobre una tecnología que aún se encuentra en fase de desarrollo. En cualquier caso, los ejemplos antes mencionados evidencian un gran interés por parte de legisladores por regular la inteligencia artificial.

Desde esta perspectiva, el *deepfake* como producto de la utilización de la inteligencia artificial, también ha estado en la agenda de los legisladores. Por un lado, al ser vectores de desinformación, pueden influir en elecciones, manipular opiniones públicas, difamar a individuos, y crear confusión sobre hechos y noticias (Chesney y Citron 2019). Además, plantean desafíos legales como la violación de derechos de imagen tanto de personas vivas como muertas y problemas de consentimiento. Lo anterior sin hablar de las posibles estafas y fraudes que se pueden realizar manipulando la imagen o voz de una persona. Así, aunque las normativas actuales sobre derechos de autor, privacidad y difamación pueden proporcionar un marco inicial para gestionar ciertos aspectos de los *deepfakes*, estas podrían resultar ocasionalmente insuficientes dada la novedad y la complejidad técnica inherente a esta tecnología (Kirchengast 2020). Esto ha sugerido la necesidad de interpretaciones legales actualizadas o nuevas leyes específicamente diseñadas para gestionar los riesgos asociados con las ultra falsificaciones. Dada la capacidad de los *deepfakes* para causar daño rápido y a gran escala, la regulación de esta tecnología puede ser necesaria.

Bajo el anterior punto de vista, ha de observarse que varios estados han incorporado en sus marcos regulatorios disposiciones que regulan la utilización de sistemas de IA con fines de creación de *deepfakes*. En particular, la versión aprobada del Reglamento Europeo de Inteligencia Artificial[1], en su artículo 50[2]

Otra vertiente observada es la adoptada por legislaciones que privilegian enfoques exclusivamente punitivos. Este tipo de regulación es delicada, dado que la instauración de tipos penales para abordar la creación y difusión de deepfakes puede acarrear varios riesgos como la posible restricción de libertades fundamentales bajo el pretexto de combatir el uso indebido de tecnologías de manipulación de imágenes y videos. En este contexto, resulta pertinente referenciar el proyecto de ley 225 del año 2024 en Colombia3. Según el texto propuesto, el artículo 296 del Código Penal, referente a la falsedad personal, se modificaría de la siguiente manera:

establece una obligación de transparencia para los usuarios de sistemas de IA que se empleen para la creación de ultra falsificaciones. Dicha obligación abarca tres aspectos fundamentales:

- **Obligación de Divulgación General:** Se requiere que se divulgue explícitamente que el contenido ha sido generado o manipulado artificialmente.
- **Contenido Artístico y Creativo:** Debe indicarse claramente que cualquier obra artística, creativa o satírica ha sido el resultado de la generación o manipulación mediante IA.
- **Manipulación de Texto:** Los textos que hayan sido objeto de manipulación deben informar al público que su contenido es resultado de dicha manipulación.

Estas medidas están diseñadas para asegurar que los usuarios de contenido generado por IA estén plenamente informados sobre la naturaleza artificial de dicho contenido, fomentando, en teoría, una mayor transparencia y confianza en el ecosistema digital. Sin embargo, la efectividad de esta norma podría ser limitada, dadas las complejidades asociadas con la imposición de sanciones y la exigencia de cumplimiento. Cabe señalar que muchos de los *deepfakes* han sido creados por individuos inescrupulosos que operan desde el anonimato proporcionado por internet. Incluso si sus identidades fueran conocidas, resultaría sumamente desafiante para las instituciones estatales supervisar cada una de las ultra falsificaciones que se producen en la red. Dicho lo anterior, es positivo que se estén estableciendo normativas que imponen obligaciones de transparencia en el uso de este tipo de tecnologías

1 Disponible en h,ps://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf. Consultado el 17 de abril de 2024

2 Los implementadores de un sistema de IA que genere o manipule contenido de imagen, audio o video que constituya una ultra falsificación, deberán revelar que el contenido ha sido generado o manipulado artificialmente. Esta obligación no se aplicará cuando el uso esté autorizado por ley para detectar, prevenir, investigar o procesar delitos. Cuando el contenido forme parte de una obra o programa evidentemente artístico, creativo, satírico, ficticio o análogo, las obligaciones de transparencia establecidas en este párrafo se limitarán a la divulgación de la existencia de dicho contenido generado o manipulado de manera apropiada que no obstaculice la visualización o el disfrute de la obra. Los implementadores de un sistema de IA que genere o manipule texto que se publique con el propósito de informar al público sobre asuntos de interés público deberán revelar que el texto ha sido generado o manipulado artificialmente. Esta obligación no se aplicará cuando el uso esté autorizado por ley para detectar, prevenir, investigar o procesar delitos, o cuando el contenido generado por IA haya pasado por un proceso de revisión humana o control editorial y donde una persona natural o jurídica tenga la responsabilidad editorial por la publicación del contenido (traducción realizada por el autor).

Otra vertiente observada es la adoptada por legislaciones que privilegian enfoques exclusivamente punitivos. Este tipo de regulación es delicada, dado que la instauración de tipos penales para abordar la creación y difusión de deepfakes puede acarrear varios riesgos como la posible restricción de libertades fundamentales bajo el pretexto de combatir el uso indebido de tecnologías de manipulación de imágenes y videos. En este contexto, resulta pertinente referenciar el proyecto de ley 225 del año 2024 en Colombia[3] Según el texto propuesto, el artículo 296 del Código Penal, referente a la falsedad personal, se modificaría de la siguiente manera:

Aquel individuo que, con el objetivo de obtener un beneficio propio o para terceros, o de causar perjuicio, realice actos de sustitución o suplantación de identidad, o se atribuya falsamente nombre, edad, estado civil, o cualquier otra cualidad que tenga efectos jurídicos, será sancionado con una multa de quince (15) a cincuenta (50) salarios mínimos mensuales legales vigentes, a menos que la conducta constituya un delito más grave.

Cuando la falsedad personal se ejecute mediante el uso de Inteligencia Artificial, la sanción aplicable será de cincuenta y uno (51) a ochocientos (800) salarios mínimos mensuales legales vigentes, siempre que la conducta no constituya otro delito.

Esta disposición legal podría ser clasificada dentro de lo que la doctrina, denomina populismo punitivo (Matthews 2005), en tanto presupone erróneamente que la problemática asociada con la creación y publicación de *deepfakes* malintencionados puede ser resuelta exclusivamente mediante la implementación de sanciones penales. Prueba de ello, es que existen varios puntos delicados en la redacción del citado proyecto de tipo penal.

Por ejemplo, la prueba de intención para determinar el "*fin de obtener un provecho o causar daño*" es compleja cuando se involucra la inteligencia artificial, lo que requiere no solo pruebas técnicas avanzadas sino también un análisis profundo de las intenciones humanas. La amplia gama en la cuantía de las multas para delitos cometidos mediante IA podría considerarse desproporcionada, planteando cuestiones sobre la proporcionalidad de la pena. Igualmente, las excepciones para usos autorizados por la ley abren potencialmente la puerta a abusos si los límites de estos usos autorizados no son definidos. En la exposición de motivos del proyecto de ley, el legislador se centra en abordar las ultra falsificaciones generadas mediante técnicas de redes generativas adversariales, omitiendo considerar que actualmente existen otras metodologías más avanzadas y efectivas para la creación de contenido sintético falso en formatos de imagen, video, texto o sonido. Esta limitación evidencia un enfoque que podría no ser suficientemente exhaustivo o actualizado frente a la rápida evolución tecnológica en este ámbito. Por último, es importante destacar que en Colombia ya existen mecanismos jurídicos que salvaguardan el derecho a la imagen y la honra de

3 Disponible en h,p://leyes.senado.gov.co/proyectos/index.php/textos-radicados-senado/p-ley-2023-2024/3122proyecto-de-ley-225-de-2024. Consultado el 17 de abril de 2024. Preservando así un equilibrio coherente en el sistema jurídico. Teniendo en cuenta las anteriores dificultades, se analizará en el último acápite la regulación de los deepfakes desde el punto de vista de sus efectos negativos en la democracia.

las personas. En este contexto, la introducción de una nueva normativa de carácter penal dirigida a regular la creación y distribución de deepfakes podría generar un conflicto normativo. Tal situación podría derivar en una superposición de leyes que, lejos de fortalecer, podría menoscabar la efectividad del marco jurídico destinado a la protección de los derechos de los ciudadanos.

Las anteriores consideraciones dejan en evidencia que la regulación de los *deepfakes* exige un análisis cuidadoso para asegurar que las normas sean efectivas, no sean desproporcionadas y que cualquier nueva legislación complemente y no contravenga las disposiciones ya existentes,

Perspectivas de regulación de deepfakes y sus consecuencias negativas en la democracia

Teniendo en cuenta los riesgos asociados a la desinformación potenciados con sistemas de inteligencia artificial, no es una sorpresa que los estados estén buscando con urgencia la creación de un marco jurídico que regule de forma efectiva la diseminación de información falsa a través de *deepfakes,* especialmente aquellos que puedan generar daños a la democracia o a las instituciones. Sin embargo, éste no constituye un fenómeno novedoso, ya que, como lo ponen de presente varios autores, desde la aparición y viralización de las noticias falsas, diversos operadores jurídicos, de diferentes estados, intentaron crear marcos regulatorios para afrontar la desinformación (França y Costa Camarão, 2022). Sin embargo, muchas de las medidas tomadas no han tenido la efectividad esperada.

A título de ilustración, se puede citar el caso del marco regulatorio creado en Francia contra las noticias falsas. Según los profesores Deffains y Thierry (2019), con el ánimo de ofrecer una respuesta a esta problemática, el legislador francés adoptó la ley nº 2018-1202 del 22 de diciembre de 2018 relativa a la lucha contra la manipulación de la información. Esta norma legal estableció un nuevo procedimiento de medidas cautelares limitado al ámbito electoral, el cual busca sacar de circulación las noticias que propagasen desinformación. Sin embargo, aquel loable mecanismo jurídico se reveló como ineficaz en la práctica, ya que exige condiciones acumulativas difíciles, como son que la noticia falsa sea introducida en los tres meses anteriores a la elección, que apunte a alegaciones o imputaciones inexactas o engañosas de un hecho capaz de alterar los escrutinios o que estas alegaciones hayan sido difundidas de manera deliberada, artificial o automatizada y masiva. Como consecuencia de lo anterior, el primer proceso de medidas cautelares fue rechazado por el Tribunal de París (Januel y Babonneau, 2019), al no cumplir varios de los requisitos de forma y su efectividad sigue en el aire. La anterior ilustración deja de presente que, en múltiples ocasiones, la sola reglamentación se hace insuficiente para controlar la desinformación. Por tanto, para combatir este fenómeno, se requiere un esfuerzo multifacético que trascienda las medidas punitivas.

Una estrategia que podría ser efectiva es la de incluir la promoción de políticas públicas que incentiven a las plataformas digitales a implementar marcas de agua claras en los contenidos gene-

rados por inteligencia artificial (IA) que estén relacionados con elecciones o cuestiones de interés general, indicando explícitamente su origen artificial. Lo anterior, bajo un modelo de co-regulación a la imagen de lo descrito por Marsden et al. (2019), en el cual las plataformas tecnológicas, en colaboración, y bajo la atenta mirada de los reguladores gubernamentales, desarrollen estándares claros y efectivos para la detección y manejo de *deepfakes* que puedan afectar a la democracia. Los contenidos que violen las normas establecidas deberían ser removidos o ir acompañados de un aviso que permita a los usuarios conocer la naturaleza del producto. Esto no solo aumentaría la transparencia, sino que también ayudaría a los usuarios a identificar y cuestionar la autenticidad del contenido que consumen.

De otra parte, la sola regulación no es efectiva sin que se invierta en educación e innovación tecnológica para mejorar la comprensión pública sobre cómo se crean y se pueden detectar las ultra falsificaciones. También sobre cómo se ha engañado a un ciudadano o elector en temas políticos. Igualmente, a la imagen de la seguridad vial, donde las políticas no solo se enfocan en sancionar las infracciones sino también en promover el diseño seguro de vehículos y la educación de los conductores, desde la regulación sobre *deepfakes* se deberían pensar métodos para fomentar el diseño y utilización de tecnologías de manera que se minimicen los riesgos y se promueva su uso seguro y ético. La conformidad con estas prácticas debería ser fomentada a través de medidas proactivas y educativas sobre las meramente punitivas, sin descartar estas últimas en los casos más extremos.

Ahora bien, al igual que Webber (2019), es de considerarse que cualquier regulación sobre los *deepfakes* debe tener en cuenta el principio de proporcionalidad, con el ánimo de que las limitaciones a su creación no se conviertan en vectores de vulneración al derecho a la libertad de expresión de los ciudadanos. Ha de recordarse que el derecho a la libertad de expresión, aunque fundamental, no es absoluto y está sujeto a ciertas limitaciones para proteger intereses y derechos importantes (Gunatilleke, 2020). En tal sentido, con el fin de proteger de abusos, la ley puede permitir ciertas restricciones, con el ánimo de prevenir, por ejemplo, la interferencia en procesos democráticos a través de las falsificaciones. No obstante, no hay que olvidar que el debate y el disenso son esenciales para una democracia sana. Bajo tal premisa, en un acápite anterior, se observó que las ultra falsificaciones pueden utilizarse de forma satírica para proponer nuevas críticas al poder. En tal sentido, la regulación de este tipo de creaciones realizadas con IA, debe ser lo suficientemente clara, de tal suerte que se minimice el riesgo de que los gobiernos usen el pretexto de combatir la desinformación para implementar medidas represivas que podrían restringir las libertades y derechos fundamentales de forma desproporcionada, como lo registran en el caso de Bangladesh (Zaman, 2022), fortaleciendo el autoritarismo digital, al permitir a los gobiernos determinar qué se considera verdadero, lo que puede favorecer a ciertos partidos políticos para monopolizar el discurso público y suprimir voces disidentes, incluidos ciudadanos, periodistas y opositores.

Bajo este tenor, es conveniente citar el informe del Foro Económico Mundial (2024), según el cual, si no se toman medidas con precaución, los esfuerzos gubernamentales por controlar la desinformación podrían, paradójicamente, conducir a la represión derechos y libertades fundamentales bajo el pretexto de definir la "verdad", complicando aún más el panorama global de la información y la confianza pública. Por ende, ha de insistirse en que la aplicación del principio de proporcionalidad puede ayudar a evitar este tipo de situaciones, por ejemplo, poniendo de presente que el aspecto crucial que determina la aceptabilidad de la creación de ultrafalsificaciones en el discurso público radica en la claridad con la que se comunica al público la naturaleza ficticia del contenido. La omisión de este aviso convierte a los deepfakes en herramientas potencialmente dañinas, capaces de desinformar y afectar negativamente a la opinión pública y al tejido social.

Conclusión

A lo largo de este texto, se ha observado que el auge de las ultra falsificaciones representa tanto una oportunidad como una amenaza para las democracias modernas. Mientras que, por un lado, ofrecen innovaciones en el ámbito creativo y tienen potenciales aplicaciones beneficiosas en varios sectores, por otro lado, su capacidad para generar desinformación altamente realista e interferir en elecciones puede socavar seriamente la confianza en las instituciones democráticas y afectar negativamente la integridad de los procesos electorales y el debate público.

Frente a los desafíos planteados por los deepfakes y su influencia en los sistemas democráticos, la principal conclusión de este análisis es la necesidad imperativa de que operadores jurídicos y la sociedad civil trabajen conjuntamente para desarrollar estrategias regulatorias equilibradas. Estas estrategias no solo deben mitigar los riesgos asociados con las ultra falsificaciones, sino también fomentar un ambiente digital transparente en el que busque evitar la vulneración de derechos fundamentales. En este sentido, es importante la adopción de normativas que aseguren transparencia en el uso de tecnologías de inteligencia artificial y promuevan la educación digital, permitiendo así que los ciudadanos diferencien entre contenido auténtico y manipulado.

Además, es esencial adoptar medidas que restrinjan toda clase de falsificación que afecte las instituciones y que protejan la integridad de las elecciones, evitando la utilización de deepfakes para influir indebidamente en el electorado y garantizar la legitimidad y equidad de los procesos democráticos. Estas acciones son fundamentales para mantener la confianza en nuestras instituciones democráticas y asegurar que la tecnología se emplee como un activo para fortalecer la democracia, y no como un vector de desestabilización.

Finalmente, es necesario que cualquier regulación tenga en cuenta la necesidad de preservar la libertad de expresión, asegurando que las medidas adoptadas no restrinjan innecesariamente la innovación tecnológica ni los derechos fundamentales. Todo lo anterior, con el ánimo de enfrentar

los riesgos de las falsificaciones que buscan generar caos a través de la desinformación y la mentira, la opacidad y los intereses personales escondidos bajo una mascarada de naturaleza algorítmica.

Referencias

AFP Colombia. (2024, 08 de marzo). Video de la televisión francesa fue manipulado para sugerir un complot para asesinar a Macron. AFP. https://factual.afp.com/doc.afp.com.34KY79H.

Akbar, M., Suaib, M., y Hussain, M. S. (2022). Analysis of Deep-Fake Technology Impacting Digital World Credibility: A Comprehensive Literature Review. *International Journal of Computer and Information Technology, 11*(2).

Bellamy, R. (2020). Lies, Deception and Democracy. *Biblioteca della libertà, 54*(225-226), 1-22. https://doi.org/10.23827/BDL_2019_3_2

Chesney, B., y Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review, 107*(1753). https://doi.org/10.15779/Z38RV0D15J

Chitra, U., y Musco, C. (2020). [Conferencia] *Analyzing the Impact of Filter Bubbles on Social Network Polarization*. Proceedings of the 13th International Conference on Web Search and Data Mining. https://doi.org/10.1145/3336191.3371825

Commaille, J. (2012). *À quoi nous sert le droit?* Gallimard.

Crootof, R. (2019). Regulating new weapons technology. En R. T. P. Alcala, & E. T. Jensen, (eds.). *The impact of emerging technologies on the law of armed conflict* (pp. 3–26). Oxford University Press. https://doi.org/10.1093/oso/9780190915322.003.0001

Crozier, M. J., Huntington, S. P., y Watanuki, J. (1975). *La crisis de la democracia: Informe sobre la gobernabilidad de las democracias a la Comisión Trilateral.* University Press.

Deffains, N., y Thierry, J. B. (2019). Fausses nouvelles. *Répertoire IP/IT et communication,* (4).

Fallis, K., Rajtmajer, S., y Obradovich, J. (2021). Political Deepfakes Are As Credible As Other Fake News and People are Concerned. *Frontiers in Communication*, 6. https://doi.org/10.3389/fcomm.2021.681106

Farid, H. (2021). Detecting and Combating Deep Fakes. *The Journal of Intelligence, Conflict, and Warfare, 3*(3), 83–87. https://doi.org/10.21810/jicw.v3i3.2752

Foro Económico Mundial. (2024, 19 de enero). The Global Risks Report 2024. https://lc.cx/pt6iwW

Gunatilleke, G. (2020). Justifying Limitations on the Freedom of Expression. *Human Rights Review,* 22, 91-108. https://doi.org/10.1007/s12142-020-00608-8

Harris, D. (2019). Deepfakes: False Pornography Is Here and the Law Cannot ProtectYou. *Duke Law & Technology Review, 17*(1), 99-127.

Huang, K. (2023, 08 de abril). Why Pope Francis Is the Star of A.I.-Generated Photos. The New York Times. https://lc.cx/ZMPJqj

Januel, P., y Babonneau, M. (2019, 21 de mayo). Loi Fake news: première application du référé. Dalloz Actualité. https://lc.cx/s2kZcj

José Rabelo França, A. y Costa Camarão, F. (2022). Regulation of fake news: National and international regulations to fight false news. *Revista Gênero E Interdisciplinaridade, 3*(03), 264–298. https://doi.org/10.51249/gei.v3i03.826

Kirchengast, T. (2020). Deepfakes and image manipulation: criminalisation and control. *Information & Communications Technology Law*, 29, 308–323. https://doi.org/10.1080/13600834.2020.1794615.

Marsden, C., Meyer, T., y Brown, I. (2019). Platform values and democratic elections: How can the law regulate digital disinformation? *Computer Law & Security Review,* 36. https://doi.org/10.1016/j.clsr.2019.105373

Matthews, R. (2005). The myth of punitiveness. *Theoretical Criminology, 9*(2), 175-201. https://doi.org/10.1177/1362480605051639.

McKay, S., & Tenove, C. (2021). Disinformation as a Threat to Deliberative Democracy. *Political Research Quarterly, 74*(3), 703-717. https://doi.org/10.1177/1065912920938143

Mediatropy. (s.f.). How AI Deepfakes are Transforming the World of Marketing and Advertising. Mediatropy. https://lc.cx/Z0uVFw

Merriam-Webster. (s.f.). What is a Deepfake? Deepfake Meaning and Examples. https://lc.cx/3PVIIq

Moskalenko, S., & Romanova, E. (2022). Deadly Disinformation: Viral Conspiracy Theories as a Radicalization Mechanism. *The Journal of Intelligence, Conflict, and Warfare, 5*(2), 129–153.

O'donnell, G. (2007). The Perpetual Crises of Democracy. *Journal of Democracy, 18*(1), 11–5. https://doi.org/10.1353/JOD.2007.0012.

Ost, F. (2016). *À quoi sert le droit? Usages, fonctions, finalités.* Larcier.

Paterson, T., y Hanley, L. (2020). Political warfare in the digital age: cyber subversion information operations and 'deep fakes'. *Australian Journal of International Affairs, 74*(4), 439-454. https://doi.org/10.1080/10357718.2020.1734772

Pawelec, M. (2022). Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. *Digital Society, 1*(19). https://doi.org/10.1007/s44206-022-00010-6

Pulido Daza, N. J., Arce, J. C., y Silva Bohórquez, A. E. (2013). El derecho a la información en Colombia: Una aproximación al estado de la información desde el derecho y los archivos. *Equidad Desarrollo*, (19), 161-190.

Reins, L. (2019). Regulating new technologies in uncertain times: Challenges and opportunities. In L. Reins, (ed.). *Regulating new technologies in uncertain times* (pp. 19-28). Springer. https://doi.org/10.1007/978-94-6265-279-8_2

Reuters Fact Check. (2024, 21 de marzo). Macron dancing clip is altered 80s nightclub footage. Reuters. https://lc.cx/j5TSyz

Silbey, J. y Hartzog, W. (2019). The Upside of Deep Fakes. *Maryland Law Review,* 78, 960.

Somers, M. (2020, 21 de julio). Deepfakes, explained. MIT Sloan. https://lc.cx/1zdiPj

Suárez, J. (2023, 24 de octubre). Posibles audios creados por IA llegan a las elecciones regionales, ¿qué tan factible es identificarlos? Colombiacheck.

https://lc.cx/lztuT_

Vokinger, K., & Gasser, U. (2021). Regulating AI in medicine in the United States and Europe. *Nature Machine Intelligence,* 3, 738-739. https://doi.org/10.1038/s42256-02100386-z

Webber, G. (2019). Proportionality and Limitations on Freedom of Speech. *Law & Society: Public Law–Constitutional Law e Journal,* 173–192. https://doi.org/10.1093/OXFORDHB/9780198827580.013.11

Woolley, S. C., y Howard, P. N. (2018). Conclusion: Political parties, politicians, and computational propaganda. In S. C. Woolley, & P. N. Howard, (eds.). *Computational propaganda: Political parties, politicians, and political manipulation on social media* (pp. 241–248). Oxford Academic. https://doi.org/10.1093/oso/9780190931407.003.0011

Zaman, F. (2022). Mechanisms of Digital Authoritarianism: The Case of Bangladesh. *SAIS Review of International Affairs,* 42, 101 - 85. https://doi.org/10.1353/sais.2022.0012

Zech, H. (2021). Liability for AI: public policy considerations. *ERA Forum,* 22, 147–158. https://doi.org/10.1007/s12027-020-00648-0.

Autor

Efraín Fandiño López. Doctor en derecho de la Universidad Paris Cité con maestrías en sociología jurídica de la Universidad Paris 2 Panthéon-Assas y en Derecho Internacional de la Universidad Lyon 3 Jean Moulin. Efraín Fandiño es investigador en varios campos del derecho de las nuevas tecnologías, teniendo un especial enfásis en inteligencia artificial desde 2017. Es investigador independiente y hace parte del Grupo de Análisis, Contexto y Estadística (GRANCE) de la Unidad de Investigación y Acusación (UIA) de la Jurisdicción Especial para la Paz (JEP).

Declaración

Conflicto de interés
No tenemos ningún conflicto de interés que declarar.
Financiamiento
Sin ayuda financiera de partes externas a este artículo.
Nota
El artículo es original y no ha sido publicado previamente.