

XXXII Congreso Nacional de Derecho Procesal.

**Comisión 1 - Inteligencia artificial: proceso penal y derechos
fundamentales (arts. 75, inc. 19 y 22, CN).**

Ponencia: Deepfakes y el derecho penal

Autores: Adrian V. Bugvila.¹

Martin Yurec²

**La tensión entre jurisdicción nacional e internacional en el cumplimiento
de sentencias de la Corte IDH**

XXXII CNDP, Mar del Plata 2024.

**Comisión 1 - Inteligencia artificial: proceso penal y derechos
Fundamentales (arts. 75, inc. 19 y 22, CN).**

Trabajo: Ponencia.

Tema: Deepfakes y el derecho penal

Autores: Adrian V. Bugvila y Martín Yurec

Síntesis:

¹ Abogado, maestrando en derecho procesal por la Universidad Nacional de La Plata; Jefe de Trabajos Prácticos de la cátedra I de derecho procesal civil y comercial de la Facultad de Ciencias Jurídicas y Sociales de la Universidad Nacional de La Plata y de la Universidad Católica de La Plata .

² Abogado, maestrando en derecho procesal por la Universidad Nacional de La Plata, Especialista en Derecho Penal por la Universidad Torcuato Di Tella

Deepfakes y el derecho penal argentino

1.- Introduccion 2.- Algunas definiciones 3.- Vulneracion a derechos fundamentales 4.- Regulación procesal 5.- Propuestas.

1.- Introduccion

En la era digital, la tecnología avanza a pasos agigantados, presentando nuevos desafíos para el derecho procesal, no ya para lo que pudiera suponer la mente de Carnelutti ³ en 1905 si no tambien para el propio Taruffo⁴ en 1992. Uno de los fenómenos más preocupantes y complejos que ha surgido en los últimos años es el de los 'deepfakes': videos, imágenes o audios generados o manipulados mediante inteligencia artificial, que resultan increíblemente realistas y difíciles de detectar. Estas creaciones plantean interrogantes fundamentales para el derecho penal, desde la tipificación de nuevos delitos hasta la admisibilidad y la construccion de un razonamiento probatorio en los procesos judiciales⁵.

Este artículo se propone analizar las implicaciones jurídicas de los deepfakes en el ámbito del derecho penal argentino, examinando las lagunas legales existentes, las posibles respuestas normativas y los desafíos probatorios que esta tecnología presenta. Asimismo, se explorará la tensión entre la libertad de expresión y la protección de derechos individuales, buscando un equilibrio que permita abordar eficazmente esta nueva realidad sin menoscabar las garantías constitucionales

2.- Algunas definiciones:

³ CARNELUTTI, Francesco. La prova civile (1915). *Milano: Giuffrè*.

⁴ TARUFFO, M., *La prova dei fatti giuridici*, Milano (1992)

⁵ FENOLL, Jordi Nieva. La prueba de los deepfakes pornográficos: IA sobre IA. *Diario La Ley*, (2024), no 10516, p. 4.

El *Deepfake* es una acrónimo entre “*Deep learning*” (un sub campo del *machine learning* o aprendizaje automático) y *fake* (falso) consistente en una red neuronal entrenada con fotos, videos o audio de personas reales⁶.

Una red neuronal es una computadora que aprende a realizar una tarea analizando un set de datos de entrenamiento con ejemplos. Mientras mayor cantidad de datos se alimenta a la red mejor es la verosimilitud de los resultados.

Existe una gran prevalencia de las Redes Adversariales Generativas (*Generative Adversarial Networks* o GAN) en la cual se compone de dos partes: Un generador crea muestras de video, imagen o audio y un Discriminador trata de distinguir las muestras generadas de las verdaderas. El Generador y el Discriminador se “enseñan” en forma mutua e iterativa.

Recientemente la aparición de un nuevo algoritmo conocido como “diffusion” produjo mejores resultados y permitió una nueva ola de grandes modelos de imagen que generan casi cualquier imagen surrealistamente convincente.

¿Por qué son relevante los *deepfakes*?

Más allá de su valor como entretenimiento inofensivo, los *deepfakes* tienen un uso siniestro. El 96% de los *deepfakes* son pornográficos⁷, usualmente no consensuada y típicamente diseñada para humillar, lastimar o buscar algún lucro.

Los *deepfakes* también forman parte de nuestro discurso político, en la forma de “*fake news*” que han creado preocupaciones considerables en el caso de las elecciones. Son claves la facilidad de la fabricación y distribución a través de apps simples como *Deepnude* o bots de telegram, los tutoriales en línea y las redes sociales.

⁶ MEIKLE, Graham. *Deepfakes*. Cambridge. *Polity*, 2023.

⁷ CHESNEY, Bobby; CITRON, Danielle. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 2019, vol. 107, p. 1753.

Actualmente a prueba de audio es todavía fácil de hacer y convincente comparado con el video y las fotografías. Existen aplicaciones populares que generan audio falso de alta calidad. Una voz puede ser clonada con un audio de whatsapp de un minuto⁸. Si se combina esta tecnología con aplicaciones como Google Duplex que fueron diseñadas para que la inteligencia artificial haga llamadas telefónicas reemplazando al usuario humano, el engaño se vuelve extremadamente simple. La detección de audio *deepfake* ha avanzado, pero esta por detrás de la detección de imágenes⁹.

4.- Regulación procesal

El Código Procesal Penal de la Provincia de Buenos Aires establece un sistema probatorio caracterizado por su flexibilidad y amplitud, pero con límites claros para proteger las garantías constitucionales. El artículo 209 consagra el principio de libertad probatoria, permitiendo acreditar hechos y circunstancias por cualquier medio de prueba establecido en el Código, e incluso por otros medios no especificados, siempre que no vulneren garantías constitucionales o afecten el sistema institucional. Este artículo también faculta a limitar pruebas superabundantes y permite prescindir de pruebas para hechos notorios con acuerdo de las partes, promoviendo así la eficiencia procesal.

En cuanto a la valoración de la prueba, el artículo 210 exige la expresión de una convicción sincera sobre la verdad de los hechos, con un desarrollo escrito de las razones que llevan a esa convicción, aplicable a todas las etapas del procedimiento, excepto en juicios por jurados donde rige la íntima convicción. Complementando estas disposiciones, el artículo 211 establece una salvaguarda crucial al declarar ineficaz toda actividad probatoria y prueba

⁸ En 2019 un abogado actuando en un caso de responsabilidad parental ("custodia") en Inglaterra patrocinando al padre que residía en Dubai logró desacreditar exitosamente prueba de audio en que aparentemente se representaba al padre como violento y agresivo. A través de peritos que pudieron analizar el archivo original de audio lograron identificar el *deepfake* que la madre había fabricado usando un foro de internet. Ryan, P. (2020, febrero 8). "*Deepfake*" audio evidence used in UK court to discredit Dubai dad. The National. Disponible en <https://www.thenationalnews.com/uae/courts/deepfake-audio-evidence-used-in-uk-court-to-discredit-dubai-dad-1.975764>

⁹ ALMUTAIRI, Zaynab; ELGIBREEN, Hebah. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, 2022, vol. 15, no 5, p. 155.

obtenida que afecte garantías constitucionales. Este marco normativo busca equilibrar la flexibilidad en la admisión de pruebas con la protección de los derechos fundamentales y la necesidad de fundamentar las decisiones judiciales.

El artículo 173 del Código Procesal Penal establece un marco detallado para el reconocimiento de elementos probatorios, abarcando documentos, objetos y otros elementos de convicción. Este artículo subraya la importancia de un proceso de reconocimiento metódico y justo, permitiendo la exhibición de estos elementos al imputado, testigos y peritos para su identificación o informe. Destaca especialmente el procedimiento para el reconocimiento de objetos, exigiendo que se invite al testigo a describir el elemento antes de su exhibición y que, cuando sea posible, se realice una exhibición conjunta con objetos similares, lo que reduce el riesgo de identificaciones erróneas. Además, el artículo extiende estas precauciones a reconocimientos de naturaleza sensorial, como voces y sonidos, equiparándolos al procedimiento de reconocimiento de personas. Esta disposición refleja un enfoque cuidadoso y garantista en la obtención y valoración de pruebas, buscando maximizar la precisión y fiabilidad de los reconocimientos en el proceso penal.

La admisibilidad de pruebas en el sistema judicial estadounidense se rige por estándares que, en la era de los deepfakes, resultan preocupantemente laxos. La Regla Federal de Evidencia 901(a)¹⁰ establece que 'para satisfacer el requisito de autenticación o identificación de un elemento de prueba, el proponente debe producir evidencia suficiente para respaldar la conclusión de que el elemento es lo que el proponente afirma que es'. Este criterio, replicado en gran medida en las jurisdicciones estatales, establece un umbral de admisibilidad notablemente bajo que, si bien facilita la presentación de pruebas, abre la puerta a potenciales manipulaciones en la era digital.

La autenticación de material videográfico, según la práctica actual, se logra mediante la demostración de que el video representa con precisión la escena

¹⁰ Federal Rules of Evidence 901 (a) *“To Satisfy the requirement of authenticating or identifying an item of evidence, the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is”* (la traducción es propia)

registrada. Sin embargo, el sistema permite que testigos sin conocimiento directo del evento puedan autenticar el material basándose en su familiaridad con las personas o lugares representados. Esta flexibilidad, aunque diseñada para facilitar el proceso probatorio, crea vulnerabilidades significativas ante la creciente sofisticación de los deepfakes.

El bajo estándar de admisibilidad actual presenta riesgos considerables en el contexto de la tecnología de deepfake. Existe la posibilidad real de que testigos, actuando de buena fe, autenticuen inadvertidamente material falsificado, o que custodios de registros certifiquen erróneamente la autenticidad de archivos manipulados. Estos escenarios no son meras especulaciones teóricas, sino amenazas tangibles a la integridad del proceso judicial.

Para contrarrestar estas amenazas, es imperativo desarrollar estrategias robustas para cuestionar la autenticidad de las pruebas visuales. Estas estrategias deben incluir la producción de evidencia que demuestre la falsedad del contenido, el cuestionamiento riguroso del origen y la cadena de custodia del material, y la utilización de peritos especializados en análisis forense de video digital. Además, es crucial aplicar un escrutinio periodístico al material, interrogando sobre su creación, temporalidad y tecnología empleada.

Ante este panorama, surge la pregunta apremiante: ¿deberíamos elevar el estándar de admisibilidad para la autenticación de material videográfico? La respuesta, aunque compleja, parece inclinarse hacia la afirmativa. Un umbral más alto de admisibilidad podría servir como salvaguarda contra la infiltración de deepfakes en el proceso judicial, preservando así la integridad del sistema. No obstante, cualquier modificación en este sentido debe equilibrarse cuidadosamente para no obstaculizar indebidamente la presentación de evidencia legítima.

En conclusión, el advenimiento de los deepfakes exige una reevaluación urgente de nuestros estándares probatorios. El sistema judicial debe evolucionar para mantenerse a la par de los avances tecnológicos, adoptando criterios más rigurosos de autenticación y aprovechando las herramientas

forenses digitales más avanzadas. Solo a través de esta adaptación podremos salvaguardar la integridad del proceso judicial en la era digital."

Reporte de Europol, alcances delincuenciales del uso de deepfakes, detección de “deepfakes”.

El laboratorio de Innovación de Europol realizó en el año 2022 un reporte detallando el uso de la tecnología de “deepfake” y sus alcances para el derecho penal. Esta tecnología puede producir contenido que convincentemente muestre a gente diciendo o haciendo actividades que nunca hizo o bien crear individuos que nunca existieron siquiera.

Atento el informe antes mencionado el crecimiento de estas tecnologías comprenden una problemática cada vez más aguda atento al alcance de múltiples actividades criminales como acoso o humillación de personas de manera online -art. 128-, extorsiones o fraudes -168, ss., 172 y ss.-, facilitar documentación fraudulenta -art. 292 y ss.-, suplantación de identidad -encontrándose en tratamiento proyectos para incorporar un art. 139 ter al código de fondo con el tipo penal específico-, pornografía no consensuada -encontrándose dentro del capítulo de extorsiones y hallándose en discusión proyectos para la creación de tipo penal específico para la extorsión cibernética de material pornográfico-¹¹, creación de material pornográfico infantil -art. 128-, falsificación o manipulación de evidencia en investigaciones criminales -art. 255-, intromisión en mercados financieros, distribución de información falsa o manipulación de la opinión pública, apoyo de grupos terroristas, apoyo a la tensión social o polarizar la opinión política. Como podemos observar un gran número de comportamientos pueden ser subsumidos dentro de los tipos penales ya existentes en el código de fondo argentino, más aún luego de las modificaciones de la ley 26.388 de delitos informáticos la cual incorporara al código de fondo variopintas modalidades delictivas como el daño informático, fraude informático, difusión de imágenes de abuso sexual infantil, violación de

11

<https://defensoria.org.ar/entradas-cpdp/la-difusion-no-consentida-de-imagenes-intimas-en-el-codigo-penal/> -consultado el 14/6/2024-.

secretos y de la privacidad, delitos contra la seguridad pública e interrupción de las comunicaciones y falsificación de documentos electrónicos.¹²

El reporte menciona posibles impactos en procesos legales a su vez haciendo hincapié en las dificultades que pueden sobrevenir sobre las pruebas audiovisuales ya que usualmente suele confiarse en su autenticidad sea cuando son extraídas del teléfono celular del imputado, descargadas de redes sociales o recibidas de un sistema de circuito cerrado de vigilancia, sea cual fuere el origen no suele estar cuestionada. Con el surgimiento de *deepfakes* cada vez será más importante el escrutinio de las imágenes para saber si el contenido es veraz o bien fue manipulado o generado artificialmente. El riesgo crece cada vez más atento el veloz avance de la tecnología en estos aspectos, permitiendo incluso reproducir en tiempo real copias aun desde un sistema de nube.

Por otro lado, a nivel nacional se han realizado nuevos avances para la persecución de este tipo de delitos cibernéticos, recientemente el Ministerio de Seguridad de la Nación ha publicado la resolución 428/24 en consonancia con la ya dictada resolución ministerial 977/19 que aprobará el Plan Federal de Prevención de Delitos Tecnológicos y Ciberdelitos, actualizado por resolución 75/22. En dichos planes el Ministerio de Seguridad comprendió que la esencia global del ciberespacio ofrece innumerables posibilidades en vista al desarrollo humano como así las mismas oportunidades, riesgos y amenazas a usuarios individuales o colectivos como a la seguridad integral de las actividades de las naciones; reconociendo un aumento producido luego del año 2020 en la ciberdelincuencia. La resolución 428/24 se apoya en los artículos 183, 235 y 243 del Código Procesal Penal Federal respecto a las responsabilidades de las Fuerzas de Seguridad en materia de persecución de ciberdelincuencia, como así estableció que ellas desarrollarán labores preventivas en el espacio cibernético en relación a infracciones y conductas contempladas en la ley de estupefacientes, amenazas y otras formas de intimidación o coacción, a los hechos contemplados en la ley 26.388, delitos relacionados con acoso sexual,

¹² Entre otros dicha ley ha reconocido la modalidad delictiva digital en múltiples tipos penales modificando los arts. 128, 153, 153bis, 155, 157, 157 bis, 173 inc. 16, 183 segundo párrafo, 184, 197 y 255.

situaciones de acoso o violencia por motivos de género, terrorismo, lavado de dinero, entre otros. Vislumbrando un marco regulatorio para el “Ciber-Patrullaje” de hechos que cuentan con una complejidad mayor atento la ubicuidad de su realización debiendo -como siempre- respetarse la libertad de expresión constitucionalmente garantizada.

Detección de “deepfakes”.

La utilización de imágenes o bien contenido audiovisual que comprenden “deepfakes” si bien es posible detectarla manualmente debido a sus inconsistencias en muchos casos es una difícil labor debiéndose chequear un limitado número de archivos como requiriendo un entrenamiento apropiado para detectar las señales. De más está decir que el proceso se ve complejizado por la predisposición humana a creer en el contenido audiovisual y tenerlo por cierto desde un punto de vista general.¹³ Los modelos generados reproducen imágenes muy convincentes pero que contienen imperfecciones detectables con una observación cercana, incluyendo el borronero de contornos alrededor de la cara, la falta de parpadeos, el reflejo en los ojos, inconsistencias en pelo, venas o cicatrices, como inconsistencias en los fondos, imágenes desenfocadas en relación al fondo, etc.

Sin embargo la detección puede ser realizada por sistemas de escaneo que generan un reporte de la autenticidad del material, las cuales son creadas con tecnología de inteligencia artificial, entrenadas usando sistemas de “*machine learning*” con bases de datos de imágenes “*deepfake*”. Estos sistemas buscan distintos tipos de señales para desenmascarar si es una imagen creada, como señales biológicas -basadas en imperfecciones de piel, tonalidades dérmicas o bien flujo sanguíneo del rostro humano-, inconsistencias en la pronunciación y la dinámica de movimiento de la boca, movimientos faciales, errores de sincronización en cuadros de video.¹⁴

¹³ Ídem Nº2, cita número 32 “Levine, T.R., ‘Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection’ Journal of Language and Social Psychology, 2014, pp. 378-392., https://www.researchgate.net/publication/273593306_Truth-Default_Theory_TDT_A_Theory_of_Human_Deception_and_Deception_Detection.”

¹⁴ McAfee, ‘The Deepfakes Lab: Detecting & Defending Against Deepfakes with Advanced AI’, 2020, accessed on 10 March 2022, <https://www.mcafee.com/blogs/enterprise/securityoperations/the-deepfakes-lab-detecting-defending-against-deepfakes-with-advanced-ai>. Citado por informe Europol

Sin perjuicio de ello la detección de “*deepfakes*” mediante tecnología basada en inteligencia artificial trae aparejada obvios inconvenientes, primero comprendamos que este tipo de tecnología se configura a partir de una base de datos, una alteración en ella podría prevenir detecciones. Por otro lado, la compresión o reducción de calidad en videos podría causar problemas de pixelado haciendo difícil la detección.

5.- Conclusiones:

Como podemos advertir la normativa de fondo se ha ido actualizando, más allá que las características de los procesos continúen siendo típicamente las mismas que siglos atrás cuando siquiera podíamos imaginar si sueñan los androides con ovejas eléctricas o bien si cabe la posibilidad de ser estafados utilizando la voz de un familiar pero producida por una inteligencia artificial. Si bien el

Ahora bien, podemos advertir distintos puntos como ser la necesidad de capacitación en materia tecnológica tanto para juristas (en este caso no solo en materia de IA o deepfakes sino en tecnología a secas) como -obviamente está decir- para profesionales especializados, que deberán adquirir nuevas herramientas difíciles de costear por estas jurisdicciones como capacitarse para poder detectar de manera manual este tipo de material.

Los deepfakes tienen el potencial de afectar a todos los actores y etapas del proceso judicial, planteando desafíos sin precedentes:

1. Abogados: Deberán desarrollar nuevas competencias para detectar y rebatir evidencia falsificada, así como para proteger a sus clientes de acusaciones basadas en deepfakes.
2. Jueces: Se enfrentarán a la compleja tarea de evaluar la autenticidad de las pruebas presentadas, requiriendo posiblemente formación especializada o el apoyo de peritos en tecnología.
3. Testigos: Podrían ser víctimas de manipulación de su imagen o voz, afectando su credibilidad o sometidos a falso testimonio sin su conocimiento.

4. Jurados: Su percepción de los hechos podría verse distorsionada por evidencia falsificada, comprometiendo la integridad del veredicto.

Además, los deepfakes pueden manifestarse en el proceso judicial de diversas formas:

- a) Creados específicamente para el litigio: Como evidencia fabricada para respaldar o refutar alegaciones.
- b) Utilizados sin conocimiento de su falsedad: Cuando una parte presenta como prueba un deepfake que cree auténtico.
- c) Como parte de registros pre-existentes: Videos de archivos o registros filmográficos que han sido alterados, planteando dudas sobre la fiabilidad de las fuentes tradicionales de evidencia.

Esta multifacética presencia de los deepfakes en el ámbito judicial exige una revisión profunda de los procedimientos de admisión y valoración de pruebas, así como la implementación de nuevas salvaguardas para garantizar la integridad del proceso judicial.