Artículo de reflexión

Trilogía
CIENCIA TECNOLOGÍA SOCIEDAD

Deepfakes como actos comunicativos: desafíos éticos sobre consentimiento y privacidad*

Deepfakes as Communicative Acts: Ethical Challenges to Consent and Privacy

Tomás Balmaceda[1], Laura F. Belli[2], Diana I. Pérez[3]

Resumen

Este artículo examina las implicaciones éticas de la tecnología *deepfake*, poniendo el foco en la relación entre consentimiento y actos comunicativos. Los *deepfakes*, posibilitados por la inteligencia artificial generativa, desdibujan la frontera entre lo auténtico y lo sintético, creando imágenes y videos que parecen reales aunque representen eventos que nunca ocurrieron. Si bien estas tecnologías pueden emplearse con fines artísticos, educativos o terapéuticos legítimos, su uso en la producción de contenidos pornográficos no consentidos expone riesgos éticos y sociales de gran envergadura. A partir de distintos debates en la filosofía del lenguaje y la ética de la tecnología, el trabajo sostiene que los *deepfakes* con contenido sexual no son éticamente neutrales, sino que su circulación constituye actos comunicativos (como insultar, silenciar o humillar) que generan daños, al margen de su veracidad fáctica. Más allá de las cuestiones de autonomía individual, estas prácticas ponen en peligro derechos fundamentales como la privacidad, la imagen y la dignidad, al mismo tiempo que erosionan la confianza pública en la evidencia audiovisual. El análisis concluye proponiendo un marco ético que ubica el consentimiento informado, la finalidad legítima y el respeto por los derechos fundamentales en el centro de cualquier uso responsable de los *deepfakes*.

Palabras clave: *deepfakes* pornográficos, inteligencia artificial generativa, lenguaje denigrante, uso responsable de la IA.

Cómo referenciar

Balmaceda, T., Belli, L. F., & Pérez, D. I. (2026). *Deepfakes* como actos comunicativos: desafíos éticos sobre consentimiento y privacidad. *Trilogía Ciencia Tecnología Sociedad, 18*(38), art. e3684. https://doi.org/10.22430/21457778.3684

* Este trabajo fue elaborado en el marco del Proyecto de Unidad Ejecutora «Inteligencia artificial, estudios científicos de la mente humana, tecnología y ética. Desafíos para el siglo XXI» del Instituto de Investigaciones Filosóficas (IIF), financiado por el Consejo Nacional de Investigaciones Científicas y Técnicas (Conicet) y la Sociedad Argentina de Análisis Filosófico (Sadaf).

[1] Universidad de San Andrés, Victoria, Argentina, tbalmaceda@udesa.edu.ar
[2] Universidad de Buenos Aires, Buenos Aires, Argentina, laurafbelli@gmail.com
[3] Universidad de Buenos Aires, Buenos Aires, Argentina, dperez@filo.uba.ar

Recibido: 4 de octubre de 2025 | **Aceptado:** 30 de enero de 2026

CC BY NC SA
OPEN ACCESS

Abstract

This article examines the ethical implications of deepfake technology, focusing on the interplay between consent and communicative acts. Enabled by generative artificial intelligence, deepfakes blur the boundary between authentic and synthetic media, creating images and videos that appear real while depicting events that never occurred. Although these technologies can serve legitimate artistic, educational, or therapeutic purposes, their use in the production of non-consensual pornographic content poses serious ethical and social risks. Building on debates in the philosophy of language and technology ethics, the paper argues that sexually explicit deepfakes are not ethically neutral: their circulation constitutes communicative acts—such as insult, silencing, or humiliation—that inflict harm regardless of their factual veracity. Beyond individual autonomy, these practices endanger fundamental rights to privacy, image, and dignity, while also eroding public trust in audiovisual evidence. The analysis concludes by proposing an ethical framework that places informed consent, legitimate purpose, and respect for fundamental rights at the center of any responsible use of deepfake technology.

Keywords: pornographic deepfakes, generative artificial intelligence, derogatory language, responsible use of AI.

INTRODUCCIÓN

En los últimos años la posibilidad de generar contenidos a través de inteligencia artificial (IA), sean textos, imágenes o sonido, ha revolucionado las sociedades occidentales. Los *deepfakes* son una de las manifestaciones más sofisticadas del uso de la tecnología conocida como inteligencia artificial generativa (IAG), gracias a la cual se crean imágenes, audios o videos a partir de datos que no se corresponden con el mundo real; a diferencia de las imágenes, audios o videos a los que estamos acostumbrados que son alguna forma de registro de sonidos o estados de cosas reales. El perfeccionamiento constante de esta tecnología lleva cada vez a un número mayor de gente a tener dificultades para discriminar contenidos generados por las IAG (fotos creadas por IA de personas que nunca existieron, cuadros, música o textos hechos «en el estilo de» algún artista realmente existente, videos de figuras públicas afirmando cosas que nunca dijeron, videos sexuales falsos, en los que aparecen caras de personas que no filmaron esos contenidos, etc.) de textos, imágenes, sonidos o videos con contenidos verídicos (fotos de hechos del mundo, cuadros o música o textos creados por el artista a quien se le atribuyen, videos de personas reales realizando acciones diversas que efectivamente hicieron, etc.). Por otra parte, la mayor accesibilidad a estas herramientas, a través de aplicaciones cada vez más económicas y fáciles de usar, ha multiplicado la creación de este tipo de imágenes, videos y textos. En combinación con internet y las redes sociales, que diseminan globalmente contenidos pronto virales que atrapan la atención, estas herramientas de generación artificial de contenido están revolucionando no solo los entornos digitales sino la vida de millones de personas.

Como toda tecnología, la IAG puede ser usada tanto para propósitos nobles como para cuestiones éticamente cuestionables. de Ruiter (2021) menciona entre sus potenciales usos benéficos su utilización

para regenerar artificialmente la voz de personas que han perdido la capacidad de hablar debido a enfermedades como la esclerosis lateral amiotrófica (ELA), por ejemplo, mejorando así la calidad de vida de los pacientes y devolviéndoles parte de su autonomía. También es posible pensar en la preservación del legado de personas fallecidas, aunque existen muchas aristas para tener en cuenta, y en la implementación en espacios de enseñanzas como el Museo Dalí en Florida, Estados Unidos, que utiliza un *deepfake* del artista para guiar a los visitantes y crear una experiencia educativa inmersiva.

Este trabajo se centrará en un uso éticamente reprobable: la producción de *deepfakes* con contenido sexual.[1] A la luz de este ejemplo se repasarán algunas de las complejidades éticas que estas tecnologías presentan, con la intención de generalizar estas reflexiones a otros usos y establecer los límites éticamente adecuados con vistas a una regulación de este tipo de tecnologías. Para abordar el problema se hará una breve presentación de qué son los *deepfakes*, qué tecnología subyace a su producción y qué efectos tienen sobre los consumidores (sección 1). En el planteo de las dificultades éticas que estos contenidos generan, se dividirán las reflexiones en diversos tópicos. En primer lugar, se analizará el proceso de generación de estas imágenes y se explorará el consentimiento que se requiere para utilizar los materiales con los que la simulación es creada (sección 2). En segundo lugar, se considerarán los diversos usos que se le puede dar a estos contenidos una vez creados, enfatizando en sus usos comunicativos (sección 3). En tercer término, se hará un repaso sobre los derechos que se ven afectados o vulnerados y se propondrán algunas pautas para el uso correcto de un *deepfake*: si utiliza la imagen o voz de la persona involucrada con su permiso; si su uso es moralmente correcto (con fines educativos, artísticos o terapéuticos dejando en claro que es una reproducción sintética y no algo verídico) y si no se viola ningún derecho. Estos principios pueden extenderse a todos los usos de *deepfakes*, independientemente de su contenido.

¿QUÉ SON LOS *DEEPFAKES*?

Se conoce como *deepfake* a una de las manifestaciones más sofisticadas del uso de la tecnología conocida como IAG para crear contenidos sintéticos con un fin específico. Es el uso de técnicas avanzadas para generar contenido falsificado (que puede ser tanto visual como sonoro o su combinación) con un nivel de realismo tan elevado que puede resultar difícil para un ser humano distinguirlo del contenido original (Whittaker et al., 2020).

Su nombre surge de la combinación de «*deep learning*», un tipo de arquitectura de IA, y «*fake*», haciendo referencia directa a la tecnología subyacente que emplea este tipo de algoritmos para modificar o crear imágenes, audio y video con el fin de producir contenidos espurios. Así, los *deepfakes* no capturan el mundo real, sino que son contenidos producidos mediante herramientas digitales potenciadas por IAG (Schick, 2020). Por lo tanto, lo que se representan no es la realidad sino un contenido generado a partir de los datos procesados.

El origen de la tecnología detrás de los *deepfakes* se remonta al menos hasta 1997, cuando los informáticos Christoph Bregler, Michele Covell y Malcolm Slaney modificaron imágenes de un video de una persona hablando para mostrarla pronunciando palabras contenidas en una pista de audio

[1] Se estima que cerca del 96 % de los *deepfakes* que se encuentran *online* son pornográficos (Ajder et al., 2019).

diferente. En los años siguientes, diversos investigadores continuaron desarrollando algoritmos para el seguimiento y emparejamiento facial. En la *3rd International Conference on Face and Gesture Recognition* se presentó el algoritmo de visión artificial Active Appearance Model (AAM), que permitió hacer coincidir un modelo estadístico de la forma y la apariencia de un objeto con una nueva imagen, y que es considerado el antecedente más importante en esta tecnología (Edwards et al., 1998). En 2016 se creó la plataforma Face2Face, que permitía intercambiar en tiempo real el área de la boca de un video con la de otra persona, que un año más tarde fue una de las tecnologías que se utilizaron para el video de Obama insultando a Donald Trump (Whittaker et al., 2020).

En la actualidad sus resultados son en ocasiones muy convincentes y su creación se ha vuelto cada vez más accesible. Gracias a herramientas de código abierto, potentes modelos de IA entrenados y hardware relativamente económico, una persona con conocimientos básicos de informática puede hoy generar videos ultrarrealistas con rostros o voces falseadas.

En términos técnicos, los *deepfakes* se basan en redes neuronales profundas, especialmente en modelos conocidos como *autoencoders* y GAN (*Generative Adversarial Networks*). Los primeros permiten tomar una imagen (por ejemplo, un rostro) y codificarla en una representación comprimida que luego se reconstruye con una «decodificadora». Para hacer *face swapping*, como sucedió con el video de Obama supuestamente insultando a Donald Trump (BBC News Mundo, 2017), se entrena a un *autoencoder* para reconocer y reconstruir las expresiones faciales de dos personas diferentes. Los GAN, por su parte, trabajan con dos redes que compiten entre sí: una generadora (que crea imágenes) y una discriminadora (que intenta distinguir las falsas de las reales). Al enfrentarse, ambas mejoran y se consigue generar imágenes cada vez más realistas. El resultado son videos donde el rostro de una persona se sobrepone al de otra, mimetizando gestos, miradas, incluso expresiones sutiles.

El material con que son entrenadas estas redes neuronales profundas es central para su éxito. En el caso del video de Obama, se requirieron cientos o miles de imágenes del rostro del expresidente estadounidense y el video del cuerpo sobre el que se insertó, que en este caso fue el de un actor. Estos datos suelen obtenerse de videos de libre acceso de YouTube, redes sociales o imágenes indexadas por buscadores como Google. Luego, se usa el software especializado que divide los rostros, identifica los puntos clave (ojos, nariz, boca), alinea las caras y entrena los modelos para reproducir expresiones, iluminación y ángulos. El entrenamiento puede durar desde algunas horas hasta varios días, dependiendo de la calidad buscada y la capacidad computacional. En cuanto al hardware, hoy basta con una computadora portátil, con una moderna unidad de procesamiento gráfico (GPU, por sus siglas en inglés), como las de la serie Nvidia RTX, y suficiente memoria RAM.

Los programas y *apps* más usados varían con el tiempo, ya que es una industria en constante expansión, pero algunos de los entornos más populares para crear *deepfakes* al momento de escribir este artículo son DeepFaceLab, una herramienta de código abierto; FaceSwap, otro proyecto *open source*, con una interfaz más amigable; y Avatarify, que puede crear *deepfakes* en tiempo real durante videollamadas y uso de *apps*.

El nivel de realismo y el poder de engaño de un *deepfake* dependen de varios factores que incluyen: la cantidad y calidad de los datos de entrenamiento —cuantas más imágenes variadas del rostro se usen, mejor aprende el modelo a simular expresiones naturales—; la resolución del video original y

del falso —ya que cuanto más baja sea, más difícil es detectar imperfecciones—; el *lip sync* para lograr que los movimientos de los labios coincidan exactamente y la integración con el cuerpo y el fondo —no basta con poner un rostro—; y, finalmente, la búsqueda cuidadosa de emparejamiento de tonos de piel, iluminación, sombras y profundidad.

De acuerdo con un estudio realizado por Lewis et al. (2023), los humanos tenemos una capacidad muy limitada para identificar *deepfakes*, incluso cuando se nos advierte explícitamente sobre su presencia. Sin advertencias previas, la mayoría de las personas no detecta la presencia de un *deepfake*, mientras que si se les informa que al menos uno de cinco videos fue manipulado digitalmente, solo el 21,6 % logra identificar correctamente cuál era. Además, casi la mitad de los encuestados señalaron erróneamente al menos un video auténtico como falso, lo que evidencia un incremento en el escepticismo frente a los contenidos audiovisuales. Es un fenómeno conocido como «dividendo del mentiroso»: a medida que las personas son conscientes de la existencia de *deepfakes*, se vuelven más propensas a dudar de la autenticidad de cualquier video, incluso los genuinos (Lewis et al., 2023, p. 2). Otras investigaciones indican que características como la familiaridad con la persona representada, la autoconfianza o el uso intensivo de las redes sociales no mejoran la precisión de la detección (Köbis et al., 2021).

Es importante destacar, por lo dicho hasta aquí, que los *deepfakes* son difícilmente distinguibles de las fotos y los videos que registran situaciones y hechos mundanos, por lo que la primera reacción es aceptarlos como imágenes y sonidos que informan acerca de situaciones y personas realmente existentes. Como argumenta Crippen (2023), dado que los *deepfakes* se ven como fotografías y que las fotografías tienen una relación causal con lo fotografiado, el contenido de verdad del *deepfake* es muy fuerte. Es decir, simplemente por ver al *deepfake* como una foto se llega a creer que lo que la imagen muestra es eso. En este sentido los *deepfakes* plantean problemas éticos relativos al (pseudo)conocimiento que parecen proveer (Kerner & Risse, 2021). Y si se duda, por el dividendo del mentiroso, se duda de todas las imágenes recibidas, porque no resulta fácil discernir cuándo hay engaño y cuándo no. En síntesis, con la reiterada exposición de estas imágenes, las personas se vuelven o demasiado crédulas o demasiado desconfiadas. Y parece que poder discernir cuándo hay engaño y cuándo no, es uno de los pilares para una vida en sociedad razonable. Con esta tecnología, parece que esta capacidad está puesta en cuestión.

Yendo al caso que ocupará a este trabajo: los *deepfakes* con contenidos pornográficos se volvieron un tema conocido a partir de 2017, debido a las denuncias de diferentes actrices populares de Hollywood sobre la manipulación audiovisual (Holliday, 2021). A finales de ese año, la primera ola de *deepfakes* pornográficos tuvo un alcance global y muchos de esos videos fueron publicados en diferentes sitios web de pornografía. Uno de los más conocidos en el momento fue el publicado por un usuario de Reddit llamado «u/*deepfake*», quien subió un video en el *subreddit* r/*deepfakes* en el que se utilizaron fotografías de Gal Gadot (actriz, protagonista de la saga Wonder Woman) sobre el cuerpo de la intérprete real de material pornográfico en el que se representaba a la celebridad teniendo sexo con su hermanastro (Fletcher, 2018). Este video marcó el inicio de una discusión global sobre los riesgos éticos de la tecnología *deepfake*, no solo en términos de desinformación, sino particularmente por su uso para perjudicar y explotar a las mujeres. Luego, se multiplicaron los videos en los cuales a

través de esta tecnología (en especial el «intercambio de rostros») se muestra a mujeres famosas en situaciones sexuales explícitas, explotando la imagen de estas celebridades sin su consentimiento. Actrices como Scarlett Johansson y Emma Watson y cantantes como Katy Perry fueron víctimas de estas prácticas. Recientemente, en 2024, la artista musical Taylor Swift fue el blanco de imágenes pornográficas falsas que se supone que se originaron en *4chan* —un sitio web que les permite a los usuarios publicar contenido de manera anónima— con la tecnología de Microsoft Designer. El material luego se divulgó a través de un canal de Telegram y mediante diferentes cuentas de usuarios de la red social X. Este caso volvió a traer a los medios el problema de los usos de esta tecnología (Goujard, 2024; Montgomery, 2024).

LA CUESTIÓN DEL CONSENTIMIENTO

El consentimiento es un tema central en las investigaciones sobre las nuevas tecnologías. En la era de las redes sociales y las comunicaciones a través de la internet, a menudo se consiente en ceder tanto los derechos como el control de mucha información personal y de imágenes propias (y las de las personas cercanas). No resulta tarea fácil negarse a hacerlo: en muchas situaciones se debe renunciar a parte de la privacidad en función de acceder a servicios, plataformas o sitios web que se desean o necesitan. Al momento de aceptar los términos y condiciones, se acepta también compartir cierta información personal.

Por esta vía se encuentra, entre los usos permitidos por haber firmado los «términos y condiciones» de la aplicación en la que se suben los contenidos, algunos a los que no se les daría el consentimiento. Como la cantidad y calidad de los datos de entrenamiento son clave para lograr *deepfakes* convincentes, la manera en la que se obtienen estos materiales es de profunda importancia.

Dejando de lado por el momento las controversias que pueden suscitarse en torno a por qué puede ser objetable la creación de *deepfakes* en pornografía (van der Nagel, 2020), se trata de un caso paradigmático en el que el consentimiento (o su ausencia) resulta especialmente complejo. La creación de material pornográfico a partir de imágenes no-sexuales (como fotografías o vídeos) viola el tipo de consentimiento que la persona retratada prestó al momento de que su imagen fuera puesta a disposición de la vista de los demás (Umbach et al., 2024). El consentimiento es uno de los pilares fundamentales que permite distinguir las prácticas y producciones sexuales entre lícitas e ilícitas (y en algunos casos, entre moralmente permisibles e impermisibles). En los últimos años, las discusiones acerca de la importancia del consentimiento en torno de las diferentes prácticas sexuales tuvieron un lugar central en el debate público (Brigham et al., 2024), especialmente a partir de los reclamos de los movimientos feministas y LGTBQI+ y de movimientos como el #metoo o el #NiUnaMenos.

No parece haber desacuerdo en que constituye un acto ilegal, así como moralmente impermisible, la utilización y circulación de imágenes sexuales (fotografías o videos) obtenidas sin el consentimiento de las personas que están siendo retratadas (por ejemplo, si son filmadas por cámaras ocultas o desde lejos), o también en el caso en que esas imágenes sean producidas por esas personas, pero obtenidas de manera ilegal (por ejemplo, por medio de robo o hackeos). Ahora bien, ¿por qué no se debería pensar lo mismo del caso de los *deepfakes*? Así como el consumo o la difusión de un acto sexual

es legal/permisible si las personas involucradas consienten en que así sea, debería ser considerada ilegal/impermisible la creación de una imagen o un video de una o más personas participando en un acto sexual, si no hubo consentimiento para tal creación.

¿Por qué el uso de la imagen identificable de una persona en pornografía *deepfake* implica una violación del consentimiento? Por un lado, porque viola el derecho a la privacidad. Estas imágenes recrean situaciones que, socialmente, se reconocen como «privadas» (se trate de actos o de partes del cuerpo). Por otro lado, se puede pensar que estas imágenes son formas de difamación, en la medida en que transmiten una mentira sobre las personas que aparecen en ellas. A su vez, en el caso de la pornografía, pueden considerarse una violación al derecho a la no alteración de la identidad social, que se basa en la protección de la posibilidad de poder ejercer la autonomía sobre el modo en que se producen las relaciones sociales con los demás (Story & Jenkins, 2023), y especialmente en el caso de elegir frente a quiénes revelar los cuerpos desnudos y las preferencias sexuales. También constituyen una forma de invasión a la privacidad sexual, un derecho que incluye el derecho a la identidad, a la intimidad y la igualdad (Citron, 2019).

Si la generación del *deepfake* parte de una fotografía o imagen original sin contenido sexual y si es una imagen de acceso público luego de que la persona retratada así lo consintiera, la distribución del original no constituye una violación a la privacidad ni una forma de difamación. Pero ¿por qué la imagen simulada (que no es «real» ni muestra un acto privado) sí lo es? Porque justamente, lo que fundamenta la exigencia de consentimiento en el caso de las fotografías y de los videos es la identificación de una persona. Si el *deepfake* pornográfico muestra a una persona identificable,[2] entonces también requiere del consentimiento de esta para ser distribuida. Cuando ese consentimiento no existe (como es justamente el caso de estas producciones) los derechos de esas personas son violados.

La creación de imágenes sexuales que involucran el uso de la semejanza o las fotografías de alguien genera una demanda especial de consentimiento. Como se mencionó al inicio de esta sección, el consentimiento brindado en relación con el uso de algunas imágenes no alcanza para que estas puedan ser utilizadas en todos los escenarios posibles, en particular que puedan ser usadas como *inputs* para generar *deepfakes* acerca de quien dio el consentimiento.

Se podría argumentar que, en ciertos casos, la persona que compartió sus imágenes dio su permiso para que se hiciera cualquier uso de las mismas. Sin embargo, el estándar de consentimiento, tal y como aparece en muchos de los acuerdos de «términos y condiciones» es muy criticado (Lane et al., 2014) porque no cumple con casi ninguno de los requerimientos que las teorías sobre el consentimiento señalan como fundamentales. Por un lado, la capacidad de rechazar el acuerdo está limitada en exceso, ya que en muchos casos se trata de servicios necesarios para la vida en sociedad (desde redes sociales que se utilizan para ámbitos laborales, hasta *apps* para acceder a turnos médicos, trámites gubernamentales, etc.). Por otro lado, presentan severas deficiencias en relación con el modo en que se transmite la información acerca de los usos a los que se consiente (desde textos con lenguaje

[2] En el caso de la pornografía *deepfake*, ya sea que se trate de personas públicas o de un tipo de la mal llamada «pornografía de venganza» (que son a menudo las formas más difundidas de este formato), la «identificabilidad» es buscada.

técnico o legal, difícil de comprender, hasta documentos con decenas de páginas, en los que resulta difícil distinguir los alcances de la decisión).

Por otro lado, como ya se mencionó, el consentimiento requerido en relación con cuestiones como el sexo y/o la pornografía no funciona de la misma manera que otros tipos de consentimiento (por el estatus de tabú que este tipo de prácticas tienen en la sociedad, por la importancia del derecho a la privacidad y por los otros motivos que se mencionaron anteriormente).

USOS COMUNICATIVOS DE LOS *DEEPFAKES*

Más allá de los problemas éticos que están involucrados en el proceso de creación de los *deepfakes*, que incluyen la cuestión del consentimiento para el uso de datos e imágenes, es necesario considerar los usos que tienen estas imágenes una vez generadas.

Por supuesto, se podría pensar en un uso estrictamente privado del *deepfake*. En este caso la imagen sería como un dibujo realizado en un diario íntimo, creado solo para los ojos del autor. Este uso difícilmente podría considerarse no ético, dado que no involucra terceros. El caso de las imágenes que sexualizan personas sería similar al de una fantasía sexual, que no parece tener nada de éticamente reprobable.[3] Sin embargo, es importante resaltar que a diferencia de un dibujo en un diario íntimo guardado en un cajón en la casa de alguien, los *deepfakes* están almacenadas en la nube y/o en los discos duros de los usuarios, por lo que, incluso si no existe la intención de difundirlo, su potencial difusión por hackeo o el acceso de terceros a estos sitios es mucho más probable que en el caso de soportes no digitales.

Además de los usos estrictamente privados (si son posibles), está el uso más frecuente y problemático, que es la difusión de estos *deepfakes* por diversos medios (en general por plataformas, sitios o redes en internet). La difusión de este material se vuelve entonces un acto comunicativo. Así, los *deepfakes* pueden ser considerados imágenes ofensivas, en la medida en que generan en quien las recibe una reacción emocional, muchas veces automática, de rechazo (Baumgartner, 2018).[4] Desde este punto de vista se podría decir que el carácter ofensivo depende de factores psicológicos individuales, es decir, de la sensibilidad que lleva al receptor a reaccionar emocionalmente. Sin embargo, en casos de *deepfakes* en los que se sexualiza a una persona o se pone en boca de una figura pública una frase o se la muestra en una actitud que jamás sostendría, el efecto ya no depende exclusivamente de la sensibilidad individual, sino que el carácter ofensivo depende de una acción denigrante que está fundada en prácticas y valores sociales, independientes de los sujetos individuales. En efecto, si alguien usa un término denigrante para referirse a otra persona, el carácter ofensivo no depende de sensibilidades individuales sino de realidades estructurales de la sociedad (Anderson & Lepore, 2013) que, de acuerdo con muchos especialistas, están codificadas en el contenido semántico del término (Camp, 2013, 2018; Jeshion, 2013). Así, aunque el contenido de este tipo de imágenes no sea lingüístico, resulta interesante pensar estos actos comunicativos como imágenes en paralelo con

[3] En Öhman (2020) se puede encontrar un análisis detallado de las similitudes y diferencias entre una *deepfake* que sexualiza a alguien sin consentimiento y una fantasía sexual.
[4] Verrips (2018) identifica tres tipos de imágenes ofensivas: las relacionada con (1) sexo, (2) lo sagrado y (3) la muerte.

diversos tipos de actos de habla que incluyen términos denigrantes. Pero, dado que, a diferencia de los términos peyorativos, no se trata de palabras de un lenguaje, sino de imágenes, en lugar de tomar en cuenta las discusiones más extendidas en el ámbito de la filosofía aplicada del lenguaje que concierne a la semántica de estos términos (Anderson & Lepore, 2024; Sosa, 2018), el enfoque podría estar en la pragmática, es decir, en el tipo de acto comunicativo que se realiza al publicar estos *deepfakes* con contenidos sexuales, humillantes o que muestran actitudes o acciones que un sujeto jamás tendría.

Lepore y Stone (2018) sostienen que el carácter peyorativo de estos actos de habla no está codificado semánticamente en los términos utilizados, sino que se trata de un «tono» con el cual se emite la proferencia en cuestión, es decir es una manera de realizar el acto ilocucionario. Por su parte, Orlando y Saab (2019) consideran que se trata de un tipo peculiar y diferente de acto de habla, que bautizan con el nombre de acto «insultivo». Langton et al. (2012) identifican tres diferentes fuerzas ilocucionarias relacionadas con el discurso de odio: (i) propaganda, (ii) ataque y (iii) habla autoritaria subordinante. En el ámbito de las discusiones sobre la pornografía como un tipo de discurso de odio, también se conciben estos actos comunicativos como de silenciamiento (Langton, 1993).[5] Situando en el contexto de la fuerza del acto comunicativo la discusión acerca del uso de los *deepfakes*, se puede introducir una dimensión ética adicional a la consideración de este tipo de imágenes, dado que acciones tales como insultar, atacar o silenciar son en sí mismas acciones éticamente reprobables.

También podría resultar interesante analizar los actos comunicativos que involucran los *deepfakes* en el contexto de las discusiones sobre libertad de expresión (Howard, 2024), derecho al que apelan muchas veces las empresas tecnológicas para permitir la difusión de todo tipo de contenido. Muchas de las discusiones en este ámbito conciernen a la libertad de expresión en lo relativo a las ideas políticas y a las posibilidades artísticas. Una de las discusiones centrales sobre este tema gira en torno a la tensión entre el derecho a la libertad de expresión y el deber de no hacer daño. Se podría, entonces, examinar el caso de las imágenes denigrantes, tanto en el ámbito político como en casos relativos a la identidad étnica, religiosa o de género, bajo esta lupa. Sin duda, las imágenes denigrantes son causal de daño.

Rini y Cohen (2022) identifican diferentes maneras en las que los *deepfakes* pueden dañar a sus objetivos, incluyendo: (1) nuevas formas de objetificación con las que se viola la autonomía de una persona al involucrarla en situaciones no consentidas y fabricadas; (2) daño ilocucionario, en el que la persona es forzada a involucrarse en actos de habla que le permitan desmentir el contenido de los *deepfakes* y (3) traumas existenciales causados por el socavamiento de la identidad y los recuerdos de una persona a través de contenidos fabricados.

Un problema adicional es que no está claro qué rol deben tener las empresas productoras de IAG que permiten generar estas imágenes, ni las plataformas y redes que las difunden. Howard (2024) menciona dos posiciones contrapuestas sobre este asunto. Por un lado, se puede entender a estas empresas como empresas privadas que tienen derecho a autorregularse de la manera que les parezca más oportuna. En el caso de las redes sociales, dado que es voluntario participar de ellas, se puede

[5] Las *deepfakes* relacionadas con la sexualización de imágenes involucran mayoritariamente mujeres, por lo que esta práctica ha sido considerada una forma peculiar de abuso sexual (Tong, 2022).

argumentar que queda en manos de las decisiones individuales de cada usuario el estar expuesto a este tipo de contenido y que curar los contenidos sería una forma de limitar la autonomía de los sujetos. Asimismo, se podría argumentar que dado que estas plataformas solo son el vehículo de la expresión de las personas que allí participan, cada una de ellas es la responsable de los actos éticamente reprobables en los que incurran, liberando a las compañías tecnológicas de toda responsabilidad moral (y jurídica). Del otro lado, se puede pensar que, en tanto las redes sociales se han vuelto un foro público, deben regir para ellas las mismas leyes que rigen en cualquier otro lugar con el fin de preservar el derecho a no ser dañado, de tener derecho a réplica, etc., es decir, que deben someterse a la legislación vigente en cada país relativa a los actos difamatorios. El problema en este caso es que pueden pertenecer a diferentes países la persona que genera el *deepfake*, la que lo publica, la empresa de IAG que posibilitó la generación del *deepfake*, la empresa que gobierna la plataforma donde se publicó y la persona cuya imagen fue dañada. Es por esta razón que la gobernanza de la IA y de las redes sociales no resulta fácil de circunscribir a las legislaciones nacionales. Por otra parte, también se puede argumentar que, dados los algoritmos de recomendación que hay en estas redes, un usuario no puede elegir no someterse a la visión de contenidos de ciertos tipos que considera ofensivos, y, por lo tanto, el derecho a no ver ciertas cosas está siendo vulnerado por estas plataformas, restándole autonomía a los usuarios.

LOS DERECHOS QUE HAY QUE PRESERVAR

Los *deepfakes* pornográficos evidencian dos cuestiones centrales en la discusión de los *deepfakes* en general: por un lado la cuestión del consentimiento y, por otro, la cuestión de los usos éticamente reprobables si se los entiende como actos comunicativos. Estos problemas se deben extender más allá del caso mencionado. Además, hay otros derechos que pueden ser violados en todo tipo de *deepfake*. En efecto, por su alto grado de realismo, representan una de las amenazas más preocupantes en relación con derechos fundamentales como el derecho a la imagen, el derecho a la privacidad y el derecho a la dignidad, entre otros.

El derecho a la imagen es una manifestación del derecho a la personalidad y está vinculado con la facultad de cada individuo para controlar la difusión de su propia representación visual. En este sentido, la imagen, entendida como uno de los principales atributos de la persona, en tanto revela sus características únicas, es también un componente esencial del libre desarrollo de la personalidad. En otras palabras, como sostiene de Ruiter (2021), la imagen de una persona es parte indisociable de su personalidad.

Tradicionalmente, la protección de este derecho se encontró unida a la necesidad de contar con el consentimiento de la persona retratada; sin embargo, con el advenimiento de esta nueva tecnología, tal exigencia está siendo puesta a prueba (Medon Affonso, 2021). Los *deepfakes* presentan un desafío directo a la protección de este derecho, ya que permiten la manipulación de imágenes y videos sin el consentimiento de la persona involucrada, en muchos casos a partir de imágenes y videos cuyo uso o distribución sí fue consentido. Como se expuso, esta práctica es especialmente alarmante cuando se crean videos pornográficos falsificados con rostros de personas reales. En este caso, la vulneración del derecho a la imagen se traduce en un daño irreparable a la reputación y a la integridad de la

persona afectada y de su entorno (McNamara, 2023). Pero no solo se vulnera la esfera personal del individuo, sino que también se pone en riesgo la veracidad de la comunicación en el espacio público.

En la segunda sección se evidenció que el uso no ético de los *deepfakes* constituye también una amenaza al derecho a la privacidad, un pilar fundamental de las sociedades democráticas. En la era digital, este derecho ha cobrado una relevancia aún mayor, dado el volumen de información personal que circula en internet y la facilidad con la que puede ser utilizada sin consentimiento (Kugler & Pace, 2021). En el caso de esta tecnología, la amenaza a la privacidad se da al permitir la creación de contenidos falsos que afectan la intimidad de los individuos y dañan su reputación (como la pornografía no consentida ya analizada, la manipulación de información sobre sus vidas, el fraude o incluso el robo de identidad). Este mal uso no solo viola el derecho a la privacidad, sino que también tiene profundas consecuencias para las víctimas en su vida personal y profesional, incluyendo sus perspectivas profesionales y sus conexiones personales (Okolie, 2023).

El derecho a la dignidad, que se encuentra estrechamente relacionado con la protección contra la difamación, la explotación y la deshumanización, es un principio fundamental también amenazado en estos contextos, cuando los videos o imágenes creadas con IA son utilizadas para humillar, manipular o desacreditar a las personas. En el ámbito personal, el caso analizado de la pornografía no consentida es un ejemplo claro, en tanto genera un daño moral y emocional profundo, afectando la autoestima y el bienestar psicológico de la víctima (Gautam et al., 2024).

El desarrollo y la proliferación de los *deepfakes* generan desafíos sin precedentes para la protección de derechos fundamentales como la imagen, la privacidad y la dignidad. Aunque la tecnología en sí misma no es intrínsecamente negativa, su uso malintencionado ha demostrado tener consecuencias graves para las personas y la sociedad en su conjunto. El uso problemático de *deepfakes* se extiende más allá de la creación de contenido falso para cumplir fantasías sexuales, también habilita tácticas como el acoso, la intimidación y el silenciamiento de personas, creando un entorno digital donde el miedo y la inestabilidad prevalecen cada vez más. Estas acciones pueden obstaculizar significativamente la libertad de expresión e inhibir la participación en el discurso público, especialmente a los grupos e individuos vulnerables (Gautam et al., 2024). Por ello resulta crucial que haya una cooperación entre la sociedad, las empresas tecnológicas y los Estados para establecer mecanismos de protección efectivos, a través de un enfoque integral y multidisciplinario que busque evitar o mitigar los efectos negativos de estas tecnologías, garantizando el respeto por los derechos fundamentales en la era digital.

CONCLUSIÓN

A lo largo de este trabajo se analizó el fenómeno de los *deepfakes* como una de las manifestaciones más complejas y disruptivas de la IAG. Se observó su evolución técnica, desde sus primeros desarrollos hasta su ubicuidad y accesibilidad actual, así como el modo en que su creciente realismo y accesibilidad han multiplicado sus efectos sociales y políticos. Se examinaron en detalle sus usos más problemáticos, especialmente en el ámbito de la sexualidad no consentida y la manipulación política, mostrando

cómo esta tecnología puede vulnerar derechos fundamentales como la privacidad, la imagen y la dignidad.

También se exploraron los mecanismos técnicos que permiten la creación de *deepfakes*, los programas más utilizados y el tipo de datos requeridos, revelando que el entrenamiento de estos modelos depende, en muchos casos, de materiales obtenidos sin consentimiento. Se analizó cómo, incluso cuando el contenido generado no es auténtico, pueden provocarse daños reales: la sola sospecha de que algo ha sido manipulado es suficiente para socavar la confianza pública, habilitando fenómenos como el dividendo del mentiroso.

Sin embargo, se advierte que los *deepfakes* no son intrínsecamente malos. Su valoración ética depende del modo en que se los utiliza y del contexto en el que circulan. A partir del análisis realizado, se propone que un uso ético de esta tecnología requiere que se cumplan simultáneamente tres condiciones:

- Consentimiento informado de las personas involucradas, tanto para el uso de su imagen o voz como para el entrenamiento de los modelos y la difusión del contenido.
- Finalidad legítima y moralmente justificada, como la creación artística, la educación, la experimentación cultural o el acompañamiento terapéutico, y no el daño, la manipulación o la explotación de otros. Esto incluye que quede claro en el receptor que está siendo testigo de un material generado artificialmente, que no responde a hechos reales. Para esto es necesario incorporar en las imágenes y audios generados artificialmente algún tipo de «marca de agua» indeleble que las distinga de registros auténticos.
- Respeto por los derechos fundamentales de las personas, incluyendo su derecho a la privacidad, a la integridad de su identidad y a no ser humilladas, desinformadas o coaccionadas a través de falsificaciones sintéticas.

Como se señaló en la introducción, los usos de estas tecnologías son diversos, y muchos de ellos traen enormes beneficios para la humanidad. Un ejemplo paradigmático de uso ético es la aplicación de *deepfakes* para la reconstrucción vocal de personas con enfermedades neurodegenerativas o daños por accidente cerebrovascular (ACV), permitiéndoles recuperar la voz con sus propios tonos y matices, e incluso crear nuevas producciones artísticas. Otro caso posible es la recreación de figuras históricas en experiencias inmersivas educativas, museográficas o artísticas, siempre que se aclare explícitamente el carácter sintético del contenido y se respete la memoria y dignidad de la persona representada. En estos contextos, los *deepfakes* pueden constituir una herramienta poderosa de accesibilidad, pedagogía o expresión creativa.

También hay casos de usos de *deepfakes* cuya legitimidad es compleja. El siguiente ejemplo complejo proviene de la política. En 2020 las elecciones de la Asamblea de Delhi, el partido Bharatiya Janata Party (nacionalista hindú) utilizó *deepfakes* de su propio dirigente Manoj Tiwari para llegar a más votantes. Mediante IA, crearon videos en que Tiwari hablaba en idiomas que él no domina, como haryanvi e inglés, y enviándolos a posibles votantes por WhatsApp (Jee, 2020). Si bien se cumple con el criterio de contar con el consentimiento del involucrado, su uso no es denigrante y no se viola ningún derecho, es discutible si quedaba claro para quien lo recibía que se trataba de un video alterado digitalmente, por lo que la creencia de que dicho dirigente era un hablante competente de la

lengua del receptor podría considerarse una creencia inducida mediante engaños. Aunque se podría pensar que lo relevante para convencer a los votantes es el contenido de lo dicho, y no la lengua en la que se dice, cuando se trata de cuestiones que atañen a las identidades (éticas, religiosas, políticas, de género, etc.), el lenguaje importa.

Por todo esto, la discusión ética no debería centrarse en prohibir o permitir los *deepfakes* en abstracto, sino en construir marcos normativos, culturales y filosóficos que ayuden a discernir cuándo, cómo y para qué su uso puede ser compatible con una sociedad justa, plural y democrática. De lo contrario, se corre el riesgo de que una herramienta con enorme potencial transformador se convierta en un instrumento de control, humillación o manipulación masiva. Frente a esa amenaza, el pensamiento crítico, el compromiso ético y la acción regulatoria informada son más necesarios que nunca.

CONFLICTOS DE INTERÉS

Los autores declaran que no presenta conflicto de interés financiero, profesional o personal que pueda influir de manera inapropiada en los resultados obtenidos o las interpretaciones propuestas.

CONTRIBUCIÓN DE AUTORÍA

Los tres autores han contribuido en partes iguales a la realización de este trabajo.

REFERENCIAS

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The State of Deepfakes: Landscape, Threats, and Impact*. Deeptrace.

Anderson, L., & Lepore, E. (2013). Slurring Words. *Noûs*, *47*(1), 25-48. https://doi.org/10.1111/j.1468-0068.2010.00820.x

Anderson, L., & Lepore, E. (2024). *The Oxford Handbook of Applied Philosophy of Language*. Oxford University Press.

Baumgartner, C. (2018). Is There Such a Thing as an 'Offensive Picture'? En B. Meyer, C. Kruse & A. M. Korte (Eds.), *Taking Offense: Religion, Art, and Visual Culture in Plural Configurations* (pp. 317-339). Brill.

BBC News Mundo. (2017, 17 de julio). *El falso Barack Obama creado con inteligencia artificial capaz de hablar como si fuera el original*. https://www.bbc.com/mundo/media-40632577

Brigham, N. G., Wei, M., Kohno, T., & Redmiles, E. M. (2024, agosto). "Violation of My Body:" Perceptions of AI-Generated Non-Consensual (Intimate) Imagery. En *Proceedings of the 20th Symposium on Usable Privacy and Security (SOUPS 2024)* (pp. 373-392). Usenix Association. https://www.usenix.org/system/files/soups2024_full_proceedings_interior.pdf

Camp, E. (2013). Slurring Perspectives. *Analytic Philosophy*, *54*(3), 330-349. https://doi.org/10.1111/phib.12022

Camp, E. (2018). Slurs as Dual-Act Expressions. En D. Sosa (Ed.), *Bad Words* (pp. 29-59). Oxford University Press.

Citron, D. K. (2019). Sexual Privacy. *Yale Law Journal*, *128*(7), 1870-1960. https://www.jstor.org/stable/45098020

Crippen, M. (2023). Conceptual and Moral Ambiguities of Deepfakes: A Decidedly Old Turn. *Synthese*, *202*(1), art. 26. https://doi.org/10.1007/s11229-023-04250-y

de Ruiter, A. (2021). The Distinct Wrong of Deepfakes. *Philosophy & Technology*, *34*(4), 1311-1332. https://doi.org/10.1007/s13347-021-00459-2

Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1998). Interpreting Face Images Using Active Appearance Models. En *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 300-305). https://doi.org/10.1109/AFGR.1998.670965

Fletcher, J. (2018). Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance. *Theatre Journal*, *70*(4), 455-471. https://doi.org/10.1353/tj.2018.0097

Gautam, A., Joshi, R. K., Narula, A., & Sharma, N. (2024). Mitigating Human Rights Violations Caused by Deepfake Technology. *Library Progress International*, *44*(3), 4628-4637.

Goujard, C. (2024, 6 de febrero). *Taylor Swift deepfakes nudge EU to get real about AI*. Politico. https://www.politico.eu/article/europe-eye-fix-taylor-swift-nude-deepfake/

Holliday, C. (2021). Rewriting the Stars: Surface Tensions and Gender Troubles in the Online Media Production of Digital Deepfakes. *Convergence*, *27*(4), 899-918. https://doi.org/10.1177/13548565211029412

Howard, J. W. (2024, 19 de enero). *Freedom of Speech*. Stanford Encyclopedia of Philosophy (Spring 2024 Edition). https://plato.stanford.edu/archives/spr2024/entries/freedom-speech/

Jee, C. (2020, 19 de febrero). *An Indian politician is using deepfake technology to win new voters*. MIT Technology Review. https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/

Jeshion, R. (2013). Expressivism and the Offensiveness of Slurs. *Philosophical Perspectives*, *27*(1), 231-259. https://doi.org/10.1111/phpe.12027

Kerner, C., & Risse, M. (2021). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics*, *8*(1), 81-108. https://doi.org/10.1515/mopp-2020-0024

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11), art. 103364. https://doi.org/10.1016/j.isci.2021.103364

Kugler, M. B., & Pace, C. (2021). Deepfake Privacy: Attitudes and Regulation. *Northwestern University Law Review*, *116*(3), 611-650. https://scholarlycommons.law.northwestern.edu/nulr/vol116/iss3/1

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (Eds.). (2014). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge University Press.

Langton, R. (1993). Speech Acts and Unspeakable Acts. *Philosophy & Public Affairs*, *22*(4), 293-330. http://www.jstor.org/stable/2265469

Langton, R., Haslanger, S., & Anderson, L. (2012). Language and Race. En G. Russell & D. G. Fara (Eds.), *Routledge Companion to Philosophy of Language* (pp. 753-767). Routledge.

Lepore, E., & Stone, M. (2018). Pejorative Tone. En D. Sosa (Ed.), *Bad Words. Philosophical Perspectives on Slurs* (pp. 132-154). Oxford University Press.

Lewis, A., Vu, P., Duch, R. M., & Chowdhury, A. (2023). Deepfake detection with and without content warnings. *Royal Society Open Science*, *10*(11), art. 231214. https://doi.org/10.1098/rsos.231214

McNamara, T. S. (2023). Is It Time to Recognize the Right to an Image? *Beijing Law Review*, *14*(4), 2109-2135. https://doi.org/10.4236/blr.2023.144117

Medon Affonso, F. J. (2021). O Direito à Imagem na Era das Deep Fakes. *Revista Brasileira de Direito Civil*, *27*(1), 251-277. https://rbdcivil.ibdcivil.org.br/rbdc/article/view/438

Montgomery, B. (2024, 31 de enero). *Taylor Swift AI Images Prompt US Bill to Tackle Nonconsensual, Sexual Deepfakes*. The Guardian. https://www.theguardian.com/technology/2024/jan/30/taylor-swift-ai-deepfake-nonconsensual-sexual-images-bill

Öhman, C. (2020). Introducing the Pervert's Dilemma: A Contribution to the Critique of Deepfake Pornography. *Ethics and Information Technology*, *22*(2), 133-140. https://doi.org/10.1007/s10676-019-09522-1

Okolie, C. (2023). Artificial Intelligence-Altered Videos (Deepfakes), Image-Based Sexual Abuse, and Data Privacy Concerns. *Journal of International Women's Studies*, *25*(2), art. 11. https://vc.bridgew.edu/jiws/vol25/iss2/11

Orlando, E., & Saab, A. (2019). Términos peyorativos de grupo, estereotipos y actos de habla. *Crítica. Revista Hispanoamericana de Filosofía*, *51*(153), 31-58. https://doi.org/10.22201/iifs.18704905e.2019.1147

Rini, R., & Cohen, L. (2022). Deepfakes, Deep Harms. *Journal of Ethics and Social Philosophy*, *22*(2), 143-161. https://doi.org/10.26556/jesp.v22i2.1628

Schick, N. (2020). *Deepfakes: The Coming Infocalypse*. Hachette UK.

Sosa, D. (Ed.). (2018). *Bad Words. Philosophical Perspectives on Slurs*. Oxford University Press.

Story, D., & Jenkins, R. (2023). Deepfake Pornography and the Ethics of Non-Veridical Representations. *Philosophy & Technology*, *36*(3), art. 56. https://doi.org/10.1007/s13347-023-00657-0

Tong, S. (2022). "You Won't Believe What She Does!": An Examination into the Use of Pornographic Deepfakes as a Method of Sexual Abuse and the Legal Protections Available to Its Victims. *UNSW Law Journal Student*, *25*, 22-25. https://www5.austlii.edu.au/au/journals/UNSWLawJlStuS/2022/25.html

Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries. En *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)* (art. 79). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642382

van der Nagel, E. (2020). Verifying Images: Deepfakes, Control, and Consent. *Porn Studies*, *7*(4), 424-429. https://doi.org/10.1080/23268743.2020.1741434

Verrips, J. (2018). A Brief Anatomy of Offensive Imagery. En B. Meyer, C. Kruse & A. M. Korte (Eds.), *Taking Offense: Religion, Art, and Visual Culture in Plural Configurations* (pp. 284-316). Brill.

Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). "All Around Me Are Synthetic Faces": The Mad World of AI-Generated Media. *IT Professional*, *22*(5), 90-99. https://doi.org/10.1109/MITP.2020.2985492