Deepfake: la imagen en tiempos de la posverdad

Deepfake: images in times of post-truth

Dr. Jacob Bañuelos Capistrán
Universidad Autónoma de Baja California
jcapis@tec.mx

Recibido: 5 de marzo de 2020.
Aceptado: 27 de marzo de 2020.

Received: *March 5th 2020.*
Accepted: *March 27, 2020.*

RESUMEN

El deepfake es una especie mediática que evoluciona progresivamente como una amenaza para los dicursos del hacer creer en tiempos de la posverdad. Este estudio tiene como objetivo construir un escenario socio-mediático sobre el fenómeno del deepfake para su comprensión y análisis, en donde se propone una clasificación de géneros, casos, campos discursivos, tecnologías de producción y detección, problemáticas éticas y retos culturales que impone esta forma emergente de establecer narrativas a través de la imagen y el sonido. Como telón de fondo de este escenario está el concepto de posverdad, entendido como un fenómeno gestado en el seno de los medios digitales y que protagoniza una guerra de los discursos del poder en el seno de los medios digitales y electrónicos. El estudio revela un escenario emergente, evolutivo, cambiante, altamente dinámico y progresivo del deepfake, como instrumento político, pieza de entretenimiento, activismo y arma de acoso que plantea nuevos retos tecnológicos, éticos, legales y culturales.

Palabras-clave: *Deepfake, posverdad, ética, tecnología, géneros, campos discursivos*

ABSTRACT

This study aims at constructing a socio-mediatic scenario that will enable an analysis of the deepfake phenomenon, a true threat to discourses in time of post-truth. To do so, it proposes a classification of genres, cases, discursive fields, technologies of production and detection, ethical issues and cultural challenges that are imposed by this emergent form of constructing narratives through image and sound. The background of this scenario is set by the concept, post-truth, understood as a phenomenon that arose within the digital media, and which has fueled a war of power discourses in such media. The deepfake scenario is revealed to be emergent, evolutionary, changing, highly dynamic and progressive; further, deepfake is identified as a phenomenon that functions as political tool, piece of entertainment, means of activism and threatening weapon that poses new technological, ethical, legal and cultural challenges.

Key Words: *Deepfake, Post-truth, Ethics, Technology, Genres, Discursive fields*

INTRODUCCIÓN

El objetivo del presente estudio es crear un escenario socio-mediático que permita comprender y analizar la complejidad del *deepfake* como un fenómeno tecnológico, cultural, legal y ético. Existe una preocupación creciente por el aumento progresivo de este tipo de expresiones, realizadas con programas cada vez más accesibles y fáciles de usar, al alcance de cualquiera que quiera producirlos con fines de entretenimiento, acoso, chantaje, propagandísticos, para construir una crítica política o para generar desinformación o *fake news*.

Las principales preocupaciones provienen del ámbito de los medios de información, medios y plataformas digitales, redes sociales, los ámbitos del poder político, la legislación, la ética y la tecnología para la detección de *deepfakes,* entendidos como documentos falsos o falsificados que se hacen pasar por auténticos, y que pueden tener un amplio alcance mediático. Se trata de un fenómeno emergente en el campo del audiovisual y por lo mismo plantea problemáticas también emergentes que no habían sido tecnológicamente gestionadas, culturalmente asimiladas o legisladas.

Las preocupaciones son principalmente de tipo ético, político, legal y tecnológico, se fundan en el hecho de que los *deepfakes* minan la credibilidad de los documentos audiovisuales, principalmente videos, como medios de información o certificación de hechos, poniendo en entredicho su veracidad o generando riesgos de desinformación, difamación o chantaje.

Los deepfakes son la punta del iceberg de lo que pasará con la imagen y cualquier otro soporte de contenidos en el universo digital y sobre todo, en la lógica de la construcción de discursos falsos y mal intencionados en la laberíntica trama de las prácticas de la posverdad. Son parte de una familia que empieza a ser numerosa, dominada o acopañada por los adjetivos *deep* y *fake*: *fake news, cheapfakes, fake nudes, shallow fakes, deep learning, deep web, deepnude,* etc. Cualquier otro término que siga al adjetivo *deep* o *fake* podría pertenecer a esta familia caracterizada por la incertidumbre.

Estirando un poco más esta lógica, podríamos hablar de una sociedad *deep* y *fake,* en donde las verdades han quedado, como siempre, en lo profundo y entredicho. La gran diferencia es que la problemática sobre qué es verdad, veracidad, verosímil, real o realidad

se ha escalado un grado más con la llegada de la inteligencia artificial al mundo de la representación y no sólo a éste, sino como régimen del orden y la construcción del sentido de realidad. La proliferación de una hiperrealidad impulsada por los avances tecnológicos en todos los campos se ha acentuado y con ella la cultura del simulacro, la hiperrealidad, la *memificación,* la *avatarización* y la falsificación como discurso.

Ya lo argumentaban tres grandes filósofos que vieron nacer y evolucionar este cambio paradigmático en los discursos sobre el *hacer creer*, Jean-François Lyotard con su *La condición posmoderna: información sobre el saber* (1979), Jean Baudrillard con su *Cultura y simulacro* (1981) y Zygmunt Bauman con su *Modernidad líquida* (2002).

Los tres filósofos parten de una teoría crítica sobre la modernidad, Lyotard previó el fin de las grandes narrativas o meta-narrativas como discursos rectores sobre la verdad, erosionadas y explosionando en múltiples formas de certificación discursiva o micro-narrativas. Baudrillard asentó las bases para la comprensión de la hiperrealidad que constituye un régimen de simulación y simulacro en el seno del sistema capitalista. Y Bauman analiza y teoriza el devenir de los valores éticos, dinámicas sociales y económicas heredadas en la modernidad por el capitalismo, como un orden en el que "todo lo sólido se desvanece en el aire" como predijo Marx, y donde el sistema moral, económico, tecnológico y la sociedad de la información imponen un orden cambiante, incierto y líquido.

Esta modernidad líquida e incierta se ve claramente representada con las problemáticas que plantean los *deepfakes,* que conforman una parte significativa de la evolución tecnológica, el desarrollo de la inteligencia artificial y *deep learining* aplicadas a la creación y alteración intencionada de imágenes. Los *deepfakes* son transformaciones narrativas y discursivas hechas con imagen, texto escrito y sonido, significan un trastocamiento total de lo valores de estos soportes como documentos de certficación de realidad y veracidad.

La noción general de un *deepfake* es que una imagen ha sido manipulada digitalmente para modificar su contenido visual, audiovisual y/o sonoro para presentarla como auténtica, cambiando el rostro de un personaje en lugar de otro, o el cuerpo y/o alterando el audio o el discurso oral del mismo. Existe un número creciente de programas cada vez más sofisticados y eficientes para alterar el video y el sonido, desde aplicaciones móviles hasta sofisticados programas y métodos de inteligencia artificial. Más adelante revisamos las tecnologías más complejas del *deepfake* que usan técnicas del *deep learning* hasta el *cheapfake* o *shallow fake*, las más más cercanas a un usuario medio.

El escenario tecno-mediático actual y futuro prevé un cambio radical en los pactos de lectura y una reconfiguración total en la relación que estableceremos con los documentos, objetos, expresiones y medios digitales. Más allá de si los documentos digitales representan la realidad, la certifican o legitiman, se impone un reto cultural en donde los documentos digitales deberán ser comprendidos no sólo como representaciones de la realidad sino como productores de la misma, independientemente de si son ciertos, falsos, legítimos o verdaderos.

TEORÍA DE LA POSVERDAD Y *DEEPFAKES*

La posverdad es un metadiscurso dentro de los discursos del hacer creer la verdad, es una estrategia discursiva intencionada, que persigue establecer una idea como verdadera a partir de una manipulación de información, hechos, actos, emociones, actores y escenarios mediáticos. La posverdad, en el campo de la política, es la nueva propaganda, sólo que más sofisticada, articulando recursos de información más precisos, con un mayor conocimiento de sus públicos objetivos y víctimas.

La posverdad tiene un rango de mediación variable, se sirve de mecanismos de viralización de información y articula la intervención de líderes de opinión, celebridades, actores políticos relevantes, medios electrónicos y digitales de información, comunidades en plataformas digitales y redes sociales. Podríamos decir que existe una arquitectura mediática de la posverdad, empleada en hacer creer que documentos falsos son auténticos y los *deepfakes* engranan perfectamente en esta arquitectura.

Existen numerosas posturas sobre la posverdad, algunas críticas y algunas que ven este fenómeno como una amenaza sobre los cánones y poderes que legitiman qué es verdad y que no lo es. La posición crítica ve en la posverdad una posibilidad en contra de un orden coservador y dominante. La noción de posverdad presupone, desde esta perspectiva, que existe una verdad, esencialista, moralista y que pertenece a ciertos actores, autoridades y medios.

Para la posición crítica, la posverdad puede abrir la posibilidad de confrontar la lógica del capitalismo en un momento dado, ya que el mismo capital la ha producido, y ser vista como "pliegue" en la noción deleuziana, es decir, como la posibilidad de generar diferencias, hibridación, apropiación, adaptación y remake: "Una diferencia que no cesa de desplegarse y replegarse" (Deleuze, 1988: 42, en Carrera, 2018).

En 2016, el Diccionario Oxford declaró posverdad (*post-truth*) como palabra del año y en 2017 *fake news*. Posverdad fue definida como un adjetivo "relacionado con o denotando circunstancias en las que los hechos objetivos son menos influyentes en la formación de la opinión pública que las apelaciones a la emoción o a la creencia personal" (Oxford Lenguages, 2016). En la teoría crítica sobre la posverdad, no hay verdades absolutas y los hechos también son sujetos a discursos que intepretan la realidad y construyen nociones de verdad intencionadas.

El término posverdad ha dado ya mucho de qué hablar y teorizar, con posturas más o menos críticas, considerado un término asociado al caos, como amenaza para la democracia o como extensión de prácticas discursivas propagandísticas habituales en el campo del poder político. Autores como Stanley (2016), Harding (2017), D´Ancona (2017), Ball (2017), Ibañez (2017), Amorós (2018), McIntyre (2018), tratan la problemática desde perspectivas políticas, sociales y mediáticas, considerando contextos coyunturales y sociológicos.

La posverdad se asocia directamente a fenómenos como las *fake news* y la construcción de discursos falsos. Sin embargo, y desde un punto de vista crítico, la falsificación ha existido desde tiempo inmemorial, al igual que la propaganda. En China, *"shanzhai"* es una disciplina que forma parte de la cultura y es una filosofía creativa (Han, 2017).

En un mundo instrumentalizado por una tecnología cada vez más avanzada, a partir de la cual se puede clonar prácticamente todo, incluso la secuencia del ADN, qué podemos esperar sobre la clonación en el terreno de las representaciones en imágenes, texto y sonido. Desde una perspectiva creativa, las tesis de Han (2017) son loables y la cultura china prospera sobre esta filosofía, que no dejan de poner en entredicho las formas que tiene Occidente de afrontar las problemáticas sobre las leyes de propiedad intelectual, la conservación patrimonial o la clonación. Para Occidente la falsificación atenta contra algunos de sus pilares discursivos fundamentales; la verdad, la originalidad, la propiedad intelectual, la propia imagen, el honor y la intimidad.

Al momento de establecer los riesgos que plantea la posverdad, ya sea como propaganda, como en el caso del Brexit o las elecciones que llevaron a Donald Trump al gobierno de Estados Unidos con ayuda de Cambridge Analytica, o bien, como discurso mediático que intencionalmente busca desinformar o crear una idea falsa sobre un hecho, personaje o una persona común, es necesario considerar los daños morales y éticos que puede causar. Estos daños sólo se pueden establecer a partir de un marco moral, ético y legal determinado, que actualmente es tan líquido e incierto como la misma posverdad.

El escenario narrativo del *deepfake* transcurre por el cause de los discursos de la falsificación en la trama digital. Participa de las estrategias de falsificación, como históricamente lo han hecho todas las formas de representación documental, visual o sonora. La fotografía es un ejemplo muy claro al ser un medio que cargó con el peso de la certificación de la realidad, como documento veraz con valor legal durante más de un siglo, hasta la llegada de la tecnología digital, ante la cual estallaron estos valores.

La fotografía siempre pudo construir narrativas verosímiles de hechos falsos, como lo demuestra su historia y la historia del fotomontaje con mayores o menores intenciones realistas. Una excelente investigación histórica sobre la manipulación fotográfica con fines propagandísticos es el libro de Alain Jaubert, *Making peoaple disappear* (Jaubert, 1989). La legitimidad de la fotografía como certificación del mundo entró en

crisis hacia los años 90 del siglo XX (Bañuelos, 2008). Una crisis que ha impuesto nuevas formas de lectura de las imágenes, oscilando entre el establecimiento cultural que impone la propia representación fotográfica y la puesta en duda sobre cualquier discurso visual, sonoro o escrito.

El punto clave de la cuestión está en establecer hasta dónde la falsificación atenta contra la dignidad humana y los derechos fundamentales. Y esto, además de ser una cuestión establecida por un régimen político-económico y legal, es una cuestión de orden moral y cultural.

ESCENARIO NARRATIVO: CASOS, GÉNEROS, CAMPOS DISCURSIVOS

Los escenarios y contenidos narrativos del *deepfake* son hipermediáticos y transmedia, participan de una memoria mediática colectiva, se alimentan de datos dejados por los propios usuarios en plataformas digitales y redes sociales, en algunos casos son anónimos, se viralizan, se comparten entre pares, forman parte de una cultura popular, mediática y de una abundante y creciente iconósfera compuesta por grandes campos discursivos: política, pornografía, entretenimiento y experimentación.

La atención prestada a los *deepfakes* va en aumento progresivo desde la aparición del primer caso detectado en la plataforma reddit, protagonizado por el usuario "deepfakes" en noviembre de 2017. El perfil r/deepfakes compartía videos porno donde los rostros de celebridades eran reemplazadas en los videos originales. Igualmente se compartió un código abierto para *deep learning* en bibliotecas populares y comenzaron a proliferar *deepnudes* de este tipo en sitios para adultos. Es posible seguir una evolución histórica del *deepfake* en los reportes anuales de Deeptrace (Patrini et al., 2018 y Ajder, 2019).

Los principales campos discursivos del *deepfake* están en la esfera de la política, la pornografía, el entretenimiento, la experimentación, los efectos en medios de información y desinformación, la tecnología de producción y de detección. Son tendencias discursivas identificadas en una gama amplia de publicaciones especializadas y periodísticas como Deeptrace (Patrini et al., 2018 y Ajder, 2019), Motherboard (Vice), The New York Times, Washington Post, The Guardian, The Economist, The Times and The BBC; y en un espectro amplio de publicaciones académicas más especializadas en tecnología de producción visual mediante inteligencia artificial, terrorismo, tecnología de detección, seguridad, ética y legislación.

DEEPFAKE PORNOGRÁFICO

De acuerdo con Deeptrace (Ajder, 2019), en 2019 aumentó al doble la cantidad de *deepfakes* encontrados en Internet. De un total de 14,678 *deepfakes* encontrados en 2019 el 96% son pornográficos, con un total de 134 millones de vistas. Nueve sitios web están dedicados exclusivamente a *deepfakes* pornográficos. Ocho de cada diez sitios pornográficos aloja *deepfakes*. La totalidad de los sitios web de *deepfakes* pornográficos ataca y daña a las mujeres.

Los ataques hechos con *deepfakes* pornográficos se centran en mujeres, músicos, actores y las nacionalidades de estos personajes son británicos, surcoreanos, norteamericanos e israelíes en mayor medida (Ajder, 2019). Los deepfakes porno, no sólo atacan a celebridades como Scarlett Johansson, Gal Gadot, Taylor Swift, Maisie Williams, Jessica Alba o periodistas como Rana Ayyub o Bharatiya Janata, sino que involucran a personalidades privadas con fines de extorsión. La suma de *avatars* y *deepfakes* porno dará la posibilidad de hacer una industria de este género a la carta y aumentará los riesgos de acoso y chantaje en los próximos años.

La aplicación DeepNude es una muestra del avance de las tecnologías del *deepfake* que permite "desnudar" a una mujer a partir de un retrato. La aplicación analiza un retrato y a partir de una base de datos de 10,000 fotos de mujeres desnudas tomadas de internet, añade un cuerpo desnudo al rostro del retrato. Esta aplicación, que ha despertado advertencias legales, alerta por sus implicaciones éticas, aunque su desarrollador se escuda diciendo que si no la hacía él, alguien más lo iba a hacer (Cole, 2019).

DEEPFAKE DE ENTRETENIMIENTO

El campo discursivo más grande para los *deepfakes* es el del entretenimiento no pornográfico, ocupa el 81% de los canales de YouTube dedicados a compartir *deepfakes*, 12% son de políticos, 5% informativos y medios de comunicación y 2% empresarios (Ajder, 2019). Existen 14 canales de YouTube dedicados a *deepfakes* y el 73% están en Estados Unidos de Norteamérica. El campo del entretenimiento está protagonizado principalmente por celebridades mediáticas populares procedentes de la cultura del espectáculo, actores de cine y personajes de películas de superhéroes.

Canales de YouTube notables de *deepfake* por su calidad de producción como *Ctrl Shift Face* o *TheFakening*, intervienen videos intercambiando rostros, cuerpos y voces de personajes famosos como Tom Cruise, Arnold Schwarzenegger, Keanu Reeves, Jennifer Lawrence, Emma Stone, Nicolas Cage, Jim Carrey, Leonardo DiCaprio, Sylvester Stallone, empresarios como Jeff Bezos, Elon Musk o políticos como Donald Trump en escenas de filmes o intercambiando rostros y voces en entrevistas de programas de televisión. Aquí la narrativa es básicamente irónica, satírica, crítica, paródica, carnavalesca y persigue fines de experimentación, comerciales y de notoriedad en las redes.

DEEPFAKE POLÍTICO

El campo discursivo más propiamente mediático es el dedicado a los *deepfakes* de políticos, ya que adquiere eco en medios electrónicos y digitales. Los personajes más relevantes en este campo son en 2018-2019 Donald Trump, Barack Obama, Vladimir Putin y Nancy Pelosi.

En 2017 la Universidad de Washington publicó un estudio sobre cómo sintentizar audio y sincronizarlo al movimiento de los labios en un video de diferente procedencia, en «Synthesizing Obama: Learning Lip Sync from Audio» (Supasorn, 2017), un antecedente tecnológico que dará pie a manipulaciones de audio en videos, como Lyrebird Ai.

En abril de 2018 BuzzFeed tuiteó un *deepfake* de Barack Obama realizado por el productor Jordan Peele con FaceApp, en el que Obama aparece diciendo que Trump es un idiota. El video se realizó con el propósito de advertir los riesgos de la desinformación que pueden generar los *deepfakes*. El post recibió más de 13K retweets y 100K vistas en YouTube en un día (Patrini, G. et al., 2018; Silverman, 2018).

En el campo político destaca la aparición de los *shallowfakes*, que a diferencia con los *deepfakes*, no requieren de grandes conocimientos en programas de *deep learning* para realizarlos. Proliferarán en todos los campos discursivos, pero son vistos como una amenaza ante las elecciones presidenciales estadounidenses de 2020 (Shao, 2020).

El *deepfake* de Nancy Pelosi es prueba de ello, lanzado en mayo de 2019, advierte también de los riesgos que un video manipulado puede originar en el marco electoral, influenciar en el electorado, dañar carreras políticas o desestabilizar el equilibrio de poder entre naciones. En este caso, la voz de Pelosi fue ralentizada para que su declaración pareciera haberla hecha en estado de ebriedad.

El *deepfake* de Pelosi alojado en el perfil de Facebook *Politics WatchDog*, recibió más de 2 millones de visitas, fue compartido 45,000 veces y generó 23,000 comentarios, mayoritariamente negativos, en 48 horas. YouTube removió el video y Facebook no quiso removerlo argumentando que sus políticas de publicación indicaban que los contenidos publicados no tenían que ser verdaderos (The Guardian, 2019).

Donald Trump se ha convertido en un auténtico campo de experimentación en las narrativas del *deepfake* político, desde un canal en YouTube llamado *Donald Trump Deepfakes*, hasta campañas de solidaridad como la de la agencia La Chose, un *deepfake* en donde se ve a Trump declarando el fin del Sida (AIDS) (Madge, 2019).

Asistiremos progresivamente a más casos como el de Pelosi en el terreno político y en otros campos discursivos. A las numerosas consecuencias éticas, morales y legales de los *deepfakes*, se suman las consecuencias políticas que pueden suscitar, la imediatez de la publicación, la viralización, el tiempo que duren en la red y la tardanza en descubrir que son falsos, factores que pueden originar daños al sistema político, a pro-

cesos democráticos y graves consecuencias en seguridad.

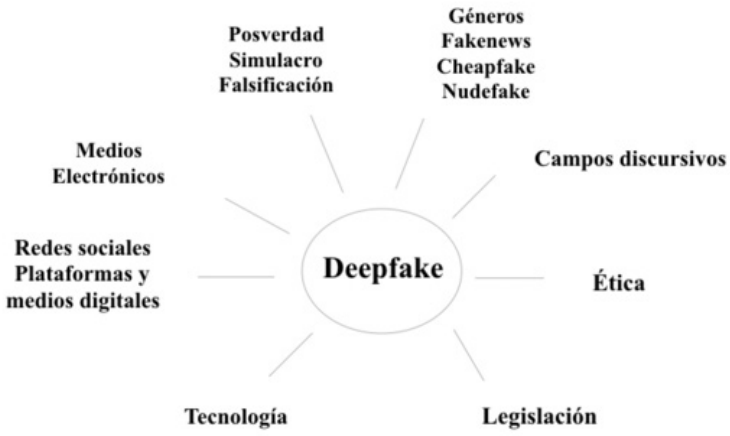
En la cultura de la desinformación, la puesta en duda de cualquier documento audiovisual es consecuencia de la normalización de la falsedad como discurso. Los *deepfakes* son un síntoma de este recorrimiento en los pactos de lectura, el estado de cosas que la digitalización, la inteligencia artificial y sus posibilidades discursivas pueden provocar.

DEEPFAKE EXPERIMENTAL

En el campo experimental destaca el artista Bill Posters quien produjo el proyecto Spectre (#Spectre) en Instagram (@bill_posters_uk) con tintes activistas, una serie de *deepfakes* realizados sobre políticos y líderes de opinión como Donald Trump, Mark Zuckerberg, Jeremy Corbyn, Boris Johnson, Kim Kardashian, en donde expresan comentarios contradictorios o provocadores. El proyecto tiene como finalidad advertir sobre los riesgos del *deepfake* como arma de desinformación (bill_posters_uk. Instagram, junio 2019).

Los géneros formales del *deepfake* emergen, se renuevan y evolucionan rápidamente. Encontramos los siguientes términos asociados, una gama que incorpora tecnología y tratamientos temáticos, programas, apps: *shallowfakes, cheapfakes, fake news, fake nudes, deepnude, face replacement, face-swap, faceshift, voice cloning, deep voice, animoji, pinscreen.*

El campo de experimentación se expande y entrecruza con los campos discursivos, géneros y tecnologías del *deepfake* en un devenir imparable y cada vez más sofisticado que pasa por la realización de rostros, avatares, sonido, voz, video conferencias, cine, fan-film, retratos animados, juegos y efectos especiales.



Esquema 1. Escenario socio-mediático del *deepfake*. Elaboración propia, 2020.

Esquema 2. Campos discursivos del *deepfake*. Elaboración propia, 2020.

TECNOLOGÍAS DEL *DEEPFAKE*

El escenario tecnológico del *deepfake* presenta dos grandes territorios: las tecnologías de producción y las de detección. Ambos campos se entrecruzan y la batalla por la detección por motivos legales o de seguridad se enfrenta a un círculo vicioso y virtuoso, mientras en un territorio se avanza, en el otro se aprende de esos avances, lo que genera el efecto de un ratón que se

muerde la cola. No habrá tecnología de detección que sea suficiente para detener la producción de *deepfakes*, sean cuales sean sus usos.

La gama tecnológica se divide igualmente en dos grandes campos: las tecnologías del *deepfake* y las del *cheapfake* (o *shallow fake*). Algunas de las tecnologías relevantes para la producción y detección de *deepfakes* se centran en el desarrollo de inteligencia artificial, aprendizaje profundo *(deep learning)*, redes neuronales convolucionales (*convolutional deep neural networks, CNN*), *machine learning*, visión computacional, reconocimiento de voz, imágenes y sonido de síntesis (Tan y Lim, 2018).

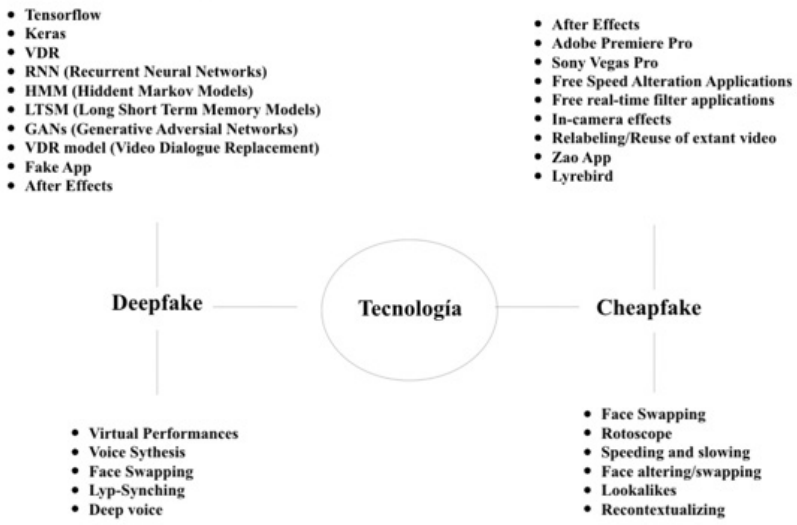
La tecnología de *modelado generativo* (*generative modelling*) a partir de *deep learning* es notable en el campo del *deepfake*, especialmente a partir de la creación del sistema GAN's (*Generative Adversarial Networks*). Inventado por Ian Goodfellow en 2014, inicia una técnica de *machine learning* en donde dos redes neuronales se confrontan y aprenden para crear nuevos datos mediante un entrenamiento. Un sistema GAN<'s (o Cycle GAN-Baekeley, 2017) entrenado en fotografías o videos existentes, genera nuevas fotografías o videos que resultan auténticos para la visión humana.

En 2016 se publica el método "Face2Face", una técnica para la recreación facial, a partir de la que es posible transferir expresiones faciales de una persona a un 'avatar' digital realista en tiempo real. En 2018 aparece FakeApp, una aplicación de escritorio construida con la librería de fuente abierta de Google, Tensorflow, lanzada con tutoriales para no expertos (Ajder, 2019).

En 2018, investigadores de la Universidad de Standford publican una investigación titulada *Deep video portraits*, un método que permite realizar animación a partir de fotos y retratos animados a partir de un solo video (Kim, et. al., 2018). DeepMind publica en 2018 el modelo Big GAN, un sistema que permite sintetizar imágenes de objetos y animales en alta resolución (Brock, A., et. al., 2018).

NVIDIA presenta "A Style-Based Generator Architecture for GANs», estableciendo un nuevo estándar para la generación de imágenes sintéticas (Karras, et. al., 2019). En 2019, la Universidad de California Berkeley publica "Everybody dance now", un método para transferir movimientos corporales de una persona de un video a otra persona en otro video (Chan, et. al. 2019).

En seguida un esquema que resume una serie de técnicas y programas de producción *deepfake* y *cheapfake* destacados entre 2014 y 2019:



Esquema 3. Tecnologías de producción *cheapkake* y *deepfake*. Basado en Paris y Donovan (2019).

TECNOLOGÍAS DE DETECCIÓN

Las tecnologías e iniciativas de detección de *deepfakes* constituyen un frente tecnológico, político, militar y empresarial urgente para la esfera política, legislativa y mediática. Un nuevo campo emerge, el *deepfake forense*. En 2016 la Agencia de Proyectos de Investigación Avanzados de Defensa (DARPA) comienza el programa Media Forensics (MediFor) para el desarrollo de métodos de análisis digital en medios. En 2018 aparece "FaceForensic" una base de datos de gran escala para entrenar herramientas de detección forense de *deepfake,* con apoyo de Technical University of Munich. Se suma a un amplio campo emergente en la investigación sobre detección de rostros y *deepfakes* (Zeinstra, 2017; Rössler, 2018).

Esfuerzos de organizaciones no gubernamentales como AI Foundation crean en 2018 un fondo de 10 millones de dóloares, para el desarrollo de herramientas que combinen técnicas de *machine learning* y moderación humana, para identificar contenidos que pudieran dañar a la gente como los *deepfakes*. Symanthec Corporation crea el mismo año un demo para detectar *deepfakes* mediante reconocimiento de rostro en BlackHat, Londres.

En 2020, Facebook, en alianza con AWS (Amazon Web Services), Microsoft, Partnership on AI's Media Integrity Steering Committee y un grupo de académicos, lanzan el Deepfake Detection Challenge para promover la innovación en técnicas de detección de *deepfakes* y contenidos mediáticos manipulados, con un premio de 1 millón de dólares (DFDC, 2020).

LEGISLACIÓN Y ÉTICA DEL *DEEPFAKE*

La evolución legal ante el *deepfake* ha sido lenta. Existen iniciativas y advertencias legales realizadas principalmente en el marco legal estadounidense. Algunas plataformas como Reddit, YouTube y Facebook comienzan a incluir en su política de publicación cláusulas que prohiben la difusión de *deepfakes*. Sin embargo, los *deepfakes* maliciosos que persiguen fines de extorsión, chantaje, acoso, o fines comerciales principalmente difundidos en sitios pornográficos, se extienden sin control legal o con anuencia de la administración de las plataformas.

La legislación sobre *deepfakes* es parte de la Ley de Autorización de Defensa Nacional para el Año Fiscal 2020 (National Defense Authorization Act for Fiscal Year 2020, NDAA) en Estados Unidos. Las iniciativas contemplan: 1) hacer un reporte de *deepfakes* usados como armas extranjeras; notificar al Congreso sobre actividades de desinformación mediante *deepfakes* en las elecciones estadounidenses; y establecer un concurso de «Deepfakes Prize» para la investigación o comercialización de tecnologías de detección de *deepfakes* (Hale, 2019).

Dos iniciativas están en curso, la *Identifying Outputs of Generative Adversarial Networks (IOGAN) Act* y la *Deepfake Report Act of 2019. En los estados de Virginia y California ya está aprobada una ley que penaliza la distribución de pornografía no consensuada mediante deepfakes. Y en Texas se penaliza la difusión deepfakes que dañen a un candidato en tiempo de elecciones* (Hale, 2019).

CONTRA *DEEPFAKES* EN MEDIOS DIGITALES

The Wall Street Journal inicia un programa de pautas para la detección de *deepfakes*. Al igual que otros medios de información, lucha contra la ola de videos falsos en las noticias, estableciendo algunas pautas para sus periodistas: examinando la fuente, encontrando versiones anteriores de las imágenes, examinando las imágenes en cámara lenta, con zoom, para detectar saltos, borrosidades, cambios de iluminación, cambios en el fondo, alteraciones del clima, voz artificial, extracción de objetos con Project Cloak (Marconi y Daldrup, 2019). De esta forma emerge un nuevo campo profesional en la industria mediática e informativa, el *media forensics* y la verificación de contenidos.

DEEPFAKES EN FACEBOOK

En 2020, Facebook cambió sus políticas de publicación sobre contenidos manipulados o *deepfakes*, con dos criterios:

» Que el contenido hayan sido editado o sintetizado, de manera que pueda ser evidente para una persona promedio, y que podría inducir a error haciendo pensar que un sujeto del video dijo palabras que no haya dicho

» Es un producto de inteligencia artificial o *machine learning* que fusiona, reemplaza o superpone contenido en un video, haciéndolo parecer auténtico.

Esta política de publicación no se aplica a contenido paródico o satírico, o a un video que ha sido editado únicamente para omitir o cambiar el orden de las palabras (Bickert, 2020).

Las nuevas políticas de publicación de Facebook son altamente ambiguas y no solucionarán el problema. A pesar de los esfuerzos de Facebook, consultando a expertos globales en tecnología, política, medios, legislación en *deepfakes*, contar con un grupo de verificadores independientes en 40 lenguas, y una asociación con socios como Partnership on AI, Cornell Tech, the University of California Berkeley, MIT, WITNESS, Microsoft, the BBC and AWS, las amenazas del *deepfake* seguirán adelante.

Para la Electronic Frontier Foundation no son necesarias nuevas leyes, se pueden aplicar las leyes existentes que ya regulan delitos asociados a los *deepfakes* maliciosos: extorsión, acoso, difamación, agravio intencional de angustia emocional (IIED), derecho de publicación y derechos de autor (Green, 2018).

Otros delitos asociados al *deepfake* son, además de extorsión y acoso, chantaje, suplantación de identidad, espionaje, violación de la privacidad, al honor, la propia imagen, difamación, daño moral, e infringir derechos de autor. La violación de los derechos humanos fundamentales y atentar contra la dignidad humana constituyen un delito y un uso anti-ético del *deepfake*, como la pornografía sin consentimiento.

Las consideraciones éticas son también ambiguas en algunos casos, por ejemplo, en los casos en que se apropia un video publicado en Internet y se usa para hacer una crítica política de un personaje público, sin ánimo de lucro ¿Es ético, es legal? No todos los *deepfakes* son ilegales o anti-éticos *per se*. Existen numerosos ejemplos en el campo experimental, artístico, político y del entretenimiento que construyen una visión crítica y satírica sobre el orden, valores y formas de poder contemporáneos.

CONCLUSIONES

El escenario socio-mediático del *deepfake* es amplio, progresivo y cambiante, se construye sobre el territorio fértil de la posverdad, que propicia la propagación de contenidos falsos mezclando hechos y emociones con fines propagandísticos o maliciosos. El *deepfake* es un elemento más de la arquitectura mediática, política, moral, ética y cultural que alberga la creación y propagación de la falsedad como producción de realidad.

La teoría crítica de la posverdad ve una posibilidad de generar discursos que cuestionen las verdades detentadas por poderes conservadores y dominantes. Sin embargo, la valoración ética y legal de los discursos de la posverdad, entre ellos los *deepfakes*, se tendrán que valorar desde su lectura y daños en un contexto moral, ético, legal y cultural específico.

La producción, difusión y distribución de *deepfakes* es imposible de detener. La lucha por detectar contenidos manipulados en medios y redes digitales no podrá detenerlos. Los *deepfakes* no deben ser entendidos como la causa de la desinformación, sino como el síntoma de una sociedad que suma los desarrollos tecnológicos a prácticas delictivas que han existido siempre. Se deberán combatir los delitos, no los *deepfakes*.

La problemática sobre la autenticidad de cualquier documento aumenta en la medida en que avanza el uso de la inteligencia artificial en la construcción de un orden material, científico y cultural. Los riesgos de la desinformación aumentan debido a la velocidad y alcance de propagación de contenidos falsos en las redes sociales, al uso de metadatos y contenidos digitales que la propia ciudadanía comparte en la red y debido a competencias de lectura sobre contenidos mediáticos heredadas del régimen mediático electrónico.

Los desarrollos tecnológicos no harán más que aumentar las posibilidades de la falsificación como discurso. Las lecciones y oportunidades que puede dejar la aparición de *deepfakes* están centradas en la necesidad de construir una nueva alfabetidad, nuevas competencias de lectura y nuevos pactos discursivos para la asimilación, confrontación, evaluación e interpretación de contenidos mediáticos digitales.

REFERENCIAS

» Ajder, H. et al. (2019). *The State of Deepfakes: Landscape, Threats, and Impact*. Amsterdam: Deeptrace.

» Amorós, M. (2019). *Fake News, la verdad de las noticias falsas*. Barcelona: Plataforma Editorial.

» Ball, J. (2017). *Post-Truth: How Bullshit Conquered the World*. London: Bite Back Publishing.

» Baudrillard, Jean (1ª ed. 1981) (1993). *Cultura y simulacro*. Barcelona: Kairós.

» Bauman, Zygmunt (1ª ed. 2002). *Modernidad líquida*. México: Fondo de Cultura Económica.

» Bickert, M. (2020). Enforcing Against Manipulated Media. Facebook. https://bit.ly/3avIBrs

» bill_posters_uk. Instagram, junio 2019. Consultado el 3 de febrero de 2020 de https://www.instagram.com/p/ByaVigGFP2U/

» Breland, A. (2019). The Bizarre and Terryfing case of the "Deepfake" that Helped Bring an African Nation to the Brink". Mother Jones, march 2019. https://bit.ly/2uX5Ewu

» Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale Gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

» Carrera, Pilar. (2018). Estratagemas de la posverdad. *Revista Latina de Comunicación Social*, 73, p. 1480. http://www.revistalatinacs.org/073paper/1317/76es.html DOI: 10.4185/RLCS-2018-1317

» Cole, S. (27 junio 2019). Esta terrorífica app crea un nude de cualquier mujer con un simple clic. *Vice*. Consultado el 3 de febrero de 2019 de https://bit.ly/2Ii3WbS

» Chan, C., et. al. (2019). Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5933-5942).

» D´Ancona, M. (2018). *Post-Truth, the New War on Truth an How to Fight Back*. United Kingdom: Edbury Press.

» Deleuze, G. (1988), *Le pli*. Paris: Les Éditions du Minuit.

» DFCD (2020). *Deepfake Detection Challenge*. https://deepfakedetectionchallenge.ai/

» Green, D. (2018). We Don't Need New Laws for Faked Videos, We Already Have Them. *Electronic Frontier Foudation*. https://bit.ly/38jWV4V

» Hale, W. (2019). First Federal Legislation on Deepfakes Signed Into Law. *JdSupra*. https://bit.ly/2wqJuDg

» Han, Byung-Chul (2017). *El arte de la falsificación y la deconstrucción en chino*. Argentina: Caja Negra Editora.

» Handing, L.(2017). *Conspiración*. Barcelona: Debate.

» Harwell, D. (2018). White House shares doctored video to support punishment of journalist Jim Acosta. *The Washington Post*. Consultado el 27 de febrero de https://wapo.st/3cytr6C

» Ibañez, J. (2017). *En la era de la posverdad*. Madrid: Calambur.

» Jaubert, A. (1989). *Making People Desappear*. New York: Pergamon-Brassey´s. International Defense Publishers, Inc.

» Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4401-4410).

» Kim, H., et. al. (2018). Deep video portraits. *ACM Transactions on Graphics (TOG), 37*(4), 1-14.

» Lyotard, Jean-François (1ª ed. 1979) (2019). *La condición posmoderna: información sobre el saber*. Madrid: Ediciones Cátedra.

» Madge, J. (2019). Has Donald Trump eradicated AIDS?. Shots.net. Consultado el 19 de febrero de 2020 de https://bit.ly/38pvo29

» Marconi, T., Daldrup, T. (2019). How The Wall Street Jorunal is preparing its journalist to detect deepfakes. NiemanLab. https://bit.ly/2wxENY1

» McIntyre, L. (2018). Post-Truth. Boston. MIT Press Essential Knowledge series.

» Oxford Langage (2016). Word of the year 2016. https://languages.oup.com/word-of-the-year/2016/

» Paris, B; Donovan, J. (2019). *Deepfakes and Cheap Fakes*. United States of America: Data & Society. https://bit.ly/2wyWHcX

» Patrini, G. et al. (2018). *The state of deepfakes: reality under attack*. Annual Report v.2.3. Amsterdam: Deeptrace.

» Rössler, A., et. al. (2018). Face forensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*.

» Shao, G. (2020). Fake videos could be the next big problem in the 2020 election. CNBC. https://cnb.cx/2VOrttc

» Silverman, C. (abril 17, 2018). How To Spot A Deepfake Like The Barack Obama–Jordan Peele Video. *BuzzFeed*. https://bzfd.it/2uQyQoR

» Supasorn, S. et al. (2017). Synthesizing Obama:Learning Lip Sync from Audio. *ACM Transactions on Graphics*, Vol. 36, No. 4, Article 95. Publication date: July 2017. https://grail.cs.washington.edu/projects/AudioToObama/siggraph17_obama.pdf

» Stanley, J. (2016). *How Propaganda Works*. New Jersey: Princenton University Press.

» Tan, K., & Lim, B. (2018). The artificial intelligence renaissance: Deep learning and the road to human-Level machine intelligence. *APSIPA Transactions on Signal and Information Processing, 7*, E6. doi:10.1017/ATSIP.2018.6

» The Guardian (2019). Real v Fake: debunking the ´drunk´Nancy Pelosi footage – video. Consultado el 28 de febrero 2020 de https://bit.ly/2VJ6IPq / https://bit.ly/3at5ADQ

» Zeinstra, C. G. (2017). Forensic Face Recognition: From characteristic descriptors to strength of evidence.