

LA REGULACIÓN DE LAS DEEPFAKES (ULTRASUPLANTACIONES) EN EL REGLAMENTO DE LA UE SOBRE INTELIGENCIA ARTIFICIAL

ANTONIO ESTELLA¹
Universidad Carlos III de Madrid

Cómo citar/Citation

Estella, A. (2025).

La regulación de las *deepfakes* (ultrasu-
plantaciones) en el reglamento de la UE sobre Inteligencia Artificial.

Revista de Administración Pública, 226, 261-290.

doi:<https://doi.org/10.18042/cepc/rap.226.11>

Resumen

Este artículo analiza de manera crítica la regulación de las *deepfakes* (ultrasu-plantaciones) tal y como figura en el art. 50 del Reglamento sobre Inteligencia Artificial. El artículo comienza realizando una serie de consideraciones generales sobre el impacto de las *deepfakes* en la democracia, examina cómo repercuten, de manera

¹ Antonio Estella es profesor titular de Derecho Administrativo y acreditado como Cate- drático de Universidad en la Universidad Carlos III de Madrid, Departamento de Derecho Público del Estado. Es titular de una Cátedra Jean Monnet «ad personam» de Derecho de la Gobernanza Económica Europea en la UC3M. Este trabajo ha sido realizado con el apoyo del proyecto SOLARIS, financiado con el Programa Horizonte Europa, del que el autor es IP en la Universidad Carlos III de Madrid. Agradezco los comentarios de los dos evaluadores anónimos de este artículo. Tal y como señala uno de ellos, es importante destacar que este artículo se centra, de manera específica, en el análisis del reglamento de IA, y deja fuera la legislación (tanto comunitaria como nacional) que podría eventualmen- te colindar con la regulación que se desprende del reglamento de IA relativa a las *deepfakes*, como, por ejemplo, la normativa española en materia de protección del derecho al honor, a la intimidad personal y familiar y a la propia imagen. Soy consciente de que esta norma- tiva colindante puede afectar a esta materia, pero he preferido dejarla fuera de este análisis para no cargar excesivamente una exposición que es ya de por sí suficientemente compleja, aunque sin duda parcial.

concreta, las *deepfakes* en el debate público, para luego pasar a examinar el art. 50 del Reglamento sobre Inteligencia Artificial en su conjunto. El resultado del análisis arroja una conclusión clara: las *deepfakes* están reguladas de manera muy defectuosa en el Reglamento sobre IA. El artículo analiza las implicaciones prácticas que se derivan de esta defectuosa regulación, para concluir, finalmente, con una serie de consideraciones de *lege ferenda*, la más importante de las cuales es la idea o concepto del estándar contrafactual.

Palabras clave

Deepfakes; ultrasuplantaciones; Inteligencia Artificial; Reglamento sobre Inteligencia Artificial; democracia; debate público; esfera pública; Administración pública; tribunales de justicia; estándar contrafactual.

Abstract

This article critically examines the regulation of *deepfakes* under Article 50 of the Regulation on Artificial Intelligence. The paper begins with a general reflection on the impact of *deepfakes* on democracy, explores how *deepfakes* specifically affect public debate, and then proceeds to analyze Article 50 of the Regulation on Artificial Intelligence in a comprehensive way. The outcome of the analysis leads to a clear conclusion: *deepfakes* are poorly regulated in the AI Regulation. The article discusses the practical implications of this flawed regulation and concludes with a series of *de lege ferenda* considerations, the most significant of which is the concept of the counterfactual standard.

Keywords

Deepfakes; Artificial Intelligence; Artificial Intelligence Regulation; democracy; public debate; public sphere; public administration; courts of justice; counterfactual standard.

SUMARIO

I. INTRODUCCIÓN. II. LA MECÁNICA DE LAS DEEPFAKES. III. LA REGULACIÓN DE LOS SISTEMAS DE IA QUE NO ESTÁN PROHIBIDOS Y QUE NO CONSTITUYEN RIESGOS ALTOS; EN PARTICULAR, LA REGULACIÓN DE LAS DEEPFAKES: 1. Sistemas de IA destinados a interactuar directamente con personas físicas. 2. Sistemas de IA de uso general que generen contenido sintético de audio, imagen, video, texto. 3. Sistemas de reconocimiento de emociones o de categorización biométrica. 4. Las *deepfakes*, en particular: sistemas de IA que generen o manipulen imágenes o contenidos de audio o video que constituyan una ultrasuplantación. 5. Sistemas de IA que generen o manipulen texto que se publique con el fin de informar al público sobre asuntos de interés público. IV. LA APLICACIÓN DEL ART. 50.4. DEL REGLAMENTO DE IA. V. CONCLUSIONES Y CONSIDERACIONES «DE LEGE FERENDA».

I. INTRODUCCIÓN

Las *deepfakes*, es decir, las «ultrasuplantaciones», por emplear el término que la versión española del reglamento de Inteligencia Artificial² utiliza para traducir la palabra «*deepfakes*», han sido declaradas por parte de algunos expertos como «uno de los mayores desafíos de la democracia, con una capacidad de disrupción mucho mayor que la que han mostrado hasta la fecha otros mecanismos de desinformación»³. Sin duda alguna, las *deepfakes* son contenidos que pueden ser empleados para «hacer el bien» en lo que se ha venido a denominar

² Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024 por el que se modifican los Reglamentos (CE) 300/2008, (UE) 167/2013, (UE) 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial) (DOUE L 12.7.2024).

³ Véase European Parliament (Panel for the Future of Science and Technology) (2021), «Tackling *Deepfakes* in European Policy».

«*deepfakes for good*». Por ejemplo, a través de sistemas de Inteligencia Artificial Generativa⁴ podemos generar la imagen de Julio César explicando, a los alumnos de Historia de un instituto, por qué este personaje histórico decidió cruzar el Rubicón⁵. El estado de desarrollo de la tecnología generativa es de tal magnitud que probablemente los oyentes tendrían la impresión (falsa, naturalmente) de que es el propio Julio César el que les está hablando. Sin embargo, no es esto lo que preocupa a los expertos que están analizando la irrupción de las *deepfakes*; lo que preocupa son precisamente los usos «para hacer el mal» (es decir, los «*deepfakes for bad*») de este tipo de contenidos. El objeto de este trabajo de investigación se centra en esta perspectiva, en la de los usos «para hacer el mal», de las *deepfakes*.

Una pregunta ha servido de base para realizar este trabajo: ¿se pueden regular de manera eficaz las *deepfakes* para proteger nuestras democracias? La formulación de esta pregunta sirve, para empezar, para definir el objeto de esta investigación. Por un lado, como hemos señalado más arriba, nos interesa la capacidad de hacer el mal de las *deepfakes*, y no su capacidad para hacer el bien, que naturalmente no negamos. Sin embargo, como señalábamos antes, es importante tener en cuenta el hecho de que las *deepfakes* tienen ese doble filo, positivo y negativo, de cara a su mejor regulación. Una regulación de los malos usos de las *deepfakes* no podría ser tan rígida como para eliminar la capacidad de las *deepfakes* para, en su caso, hacer el bien. Uno de los mayores retos que tiene planteada la regulación de este fenómeno es, precisamente, este: regular para evitar y en su caso castigar los malos usos, de tal manera que no se impida el despliegue de los sistemas generadores de *deepfakes* que tengan por objeto hacer el bien. En segundo lugar, la pregunta que hemos realizado más arriba nos indica que no nos interesan los impactos negativos que las *deepfakes* puedan tener en la privacidad de las personas. En efecto: el fenómeno de las *deepfakes* surge cuando algunos desarrolladores de *deepfakes* empezaron a emplear sistemas de IA para insertar la cara de actrices conocidas (como, por ejemplo, Scar-

⁴ La OCDE define los sistemas de IA generativa como «una categoría de IA que crea nuevo contenido como texto, imágenes, video y música» (véase aquí: <https://is.gd/xz1hsT>. (LLMs). En el Informe «Initial Policy Considerations for Generative AI» (OECD Artificial Intelligence Papers, September 2023, 1), la OCDE especifica esta definición diciendo que «los sistemas de inteligencia artificial generativa crean nuevo contenido —incluyendo texto, imágenes, audio y video— basándose en sus datos de entrenamiento y en respuesta a instrucciones». La IA generativa es especialmente relevante a los efectos de este artículo porque son los sistemas de IA generativa los que se emplean para generar o manipular *deepfakes*. Hay que recordar que el reglamento de Inteligencia Artificial no regula de manera directa la IA generativa. Esto es criticado por Wachter (2024).

⁵ Véase aquí: (<https://is.gd/6IQB3c>) esta recreación de la imagen de Julio César por IA generativa usando información de los bustos de Julio César de su época. Sobre la cuestión de por qué César tomó la decisión de cruzar el Rubicón, véase, por ejemplo, Stanton (2003).

lett Johansson) en videos de contenido pornográfico⁶. Este uso negativo de las *deepfakes* tiene un impacto puramente privado que queda fuera, por tanto, del objeto de este trabajo. Nos interesan, sin embargo, los impactos «públicos» que puedan tener los malos usos de las *deepfakes*, y en concreto, en la democracia. Cómo proteger la democracia de las *deepfakes* es, por tanto, la cuestión fundamental que se plantea este artículo⁷.

Pongamos en contexto esta problemática. Lo primero que hay que señalar es que, aunque a nosotros, ciudadanos de Estados miembros de la Unión Europea, la democracia nos parezca un sistema político connatural a nuestras sociedades, solamente un 29% de la población mundial vive en sistemas democráticos (datos de 2023) (unos 2.300 millones de personas). Si examinamos la cuestión desde un punto de vista dinámico, vemos que la situación es todavía más preocupante, puesto que la democracia está en declive, es decir, va habiendo cada vez más países que pasan de ser democracias a ser regímenes no-democráticos. Concretamente, en 2003 un 50% de la población vivía en regímenes autocráticos, mientras que en 2023 este porcentaje se ha elevado al 71% (unos 5.700 millones de personas) (véase gráfico 1). No es una exageración decir que la democracia está amenazada en nuestro planeta⁸.

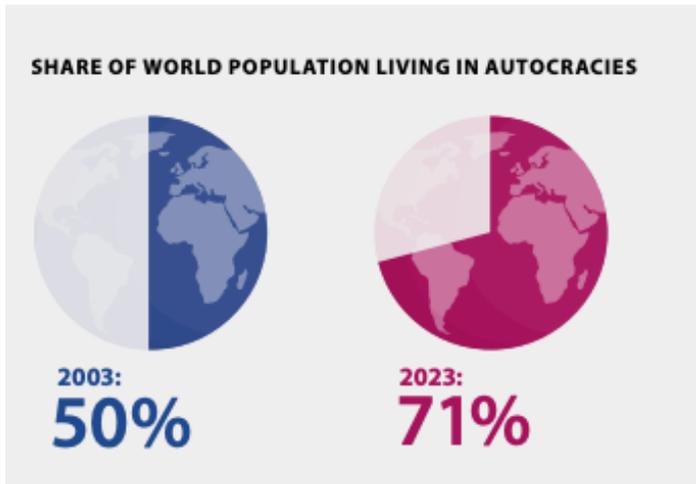
En este contexto repercuten de manera especialmente negativa los fenómenos de desinformación, de noticias falsas, y, finalmente, las ultrasuplantaciones, como corolario de un proceso que se lleva gestando desde hace ya unos cuantos años. Si bien el foco de este trabajo está en estas últimas, las *deepfakes*, no podemos perder de vista que el panorama en el que las *deepfakes* están actuando es mucho más amplio y no se limita a ellas.

⁶ Véase D. Harwell (2018), «Scarlett Johansson on fake AI-generated sex videos: “Nothing can stop someone from cutting and pasting my image”». The Washington Post, 31 December 2018. Véase también E. Meskys, J. Kalpokiene, P. Jurcys y A. Liaudanskas (2020), *Regulating Deep Fakes: Legal and Ethical Considerations*.

⁷ J. L. Piñar y M. Barrio (2023) (coords.): «IA, democracia y derechos fundamentales». *Derecho Digital e Innovación*, 18 (octubre-diciembre).

⁸ Sobre el concepto de democracia, véase Przeworski, A. (2024), «Who decides what is democratic?» *Journal of Democracy*, 35, 3, págs. 5-16. Este autor discute dos conceptos de democracia, uno como «método» para resolver disputas y otro como «la encarnación de valores, ideales, o intereses que diferentes grupos de personas quieren que la democracia lleve a cabo». Incluso aunque adoptemos una concepción minimalista de democracia, la irrupción de las *deepfakes* afecta a la democracia como mero método para resolver disputas, es decir, en una mera concepción minimalista de la democracia. Me interesa también en este sentido el concepto de democracia de Habermas como «esfera pública» [J. Habermas (1998), *Between Facts and Norms: contributions to a discourse Theory of Law and Democracy*. MIT Press].

Gráfico 1. Porcentaje de la población mundial que vive bajo regímenes autocráticos



Fuente: V-Dem (2024:7). Consultado última vez el 28 de octubre de 2024.

En este contexto, las preguntas que nos hacemos, de manera más detallada, son, pues, las siguientes: ¿qué puede hacer el derecho para contrarrestar y corregir esta tendencia? ¿Cuál es la respuesta concreta que desde el ámbito jurídico se puede estructurar en relación con el fenómeno de las *deepfakes* «para hacer el mal»? ¿Cómo puede el derecho regular las *deepfakes* «para hacer el mal» sin, al mismo tiempo, reprimir la innovación de los sistemas de IA generadores de *deepfakes* «para hacer el bien»? ¿Cuáles son los mecanismos a través de los cuales las *deepfakes* hacen daño a la democracia? ¿Cómo proteger a la democracia de las *deepfakes* sin al mismo tiempo generar una desconfianza generalizada en el sistema político, en lo que se ha venido a denominar «the liar's dividend»? ¿La regulación de las *deepfakes* para los malos usos, es una cuestión de *hard law*, o tiene cabida el *soft law* en este ámbito? ¿Quién controla y ejecuta la regulación en materia de *deepfakes* para evitar sus usos nocivos? ¿qué papel tienen las administraciones públicas y los tribunales de justicia en este ámbito?

Vamos a examinar todas estas preguntas en lo que sigue. Empezaremos examinando el mecanismo a través del cual operan las *deepfakes* «para hacer el mal» a la democracia.

II. LA MECÁNICA DE LAS DEEPFAKES

Para adentrarnos en la forma a través de la cual las *deepfakes* erosionan la democracia, tenemos que empezar dando una noción de *deepfakes*. Con carácter general, las *deepfakes* suponen la generación «ex novo» y/o la manipulación de

video, audio, texto y/o imagen a través de sistemas de Inteligencia Artificial Generativa, de tal forma que se produzca un resultado que parece verdadero cuando no lo es^{9,10}. Las palabras fundamentales de esta definición son, por tanto, «deep», que hace alusión a la creación de contenidos a través de sistemas de IA generativa, y la palabra «fakes», que hace alusión a su carácter no verdadero. La capacidad de los actuales sistemas de IA generativa para producir *deepfakes* realistas es en la actualidad muy elevada, de tal manera que en muchas ocasiones es casi imposible a la vista del ojo humano detectar cuándo un contenido es falso o es verdadero¹¹.

Por otro lado, desde un punto de vista jurídico, existen distintas definiciones legales de «*deepfake*». Para empezar, el reglamento de IA, en su art. 3.60), señala que las «ultrasuplantaciones» son «un contenido de imagen, audio o video generado o manipulado por una IA que se asemeja a personas, objetos, lugares, entidades o sucesos reales y que puede inducir a una persona a pensar erróneamente que son auténticos o verídicos». La diferencia que tiene esta definición con la que hemos dado más arriba es, fundamentalmente, que deja fuera al *texto*, lo cual no parece del todo justificado, como veremos en este trabajo. Pensemos, por ejemplo, en ChatGPT, un sistema de IA generativa que genera texto; ChatGPT tiene la capacidad de generar textos que son falsos (pongamos por caso el texto de un discurso que supuestamente pronunciaría Donald Trump y que se filtraría a la prensa diciendo que Trump apoya de forma expresa el llamado «proyecto 2025»)¹².

Una vez que hemos establecido cuál es la definición de *deepfakes* que entendemos es más adecuada, podemos pasar a examinar la cuestión de la mecánica

⁹ A. Birrer y N. Just (2024), «What we know and don't know about *deepfakes*: An investigation into the state of the research and regulatory landscape», *New Media and Society* 0(0)), nos advierten de que a pesar del incremento reciente de literatura académica sobre *deepfakes*, no hay una «definición universalmente aceptada» de *deepfakes*. De los 79 artículos doctrinales que analizan, 36 incluyen en la definición videos, 17 imágenes, y 15 audio. La inclusión de texto es menos habitual, pero sin embargo algunos autores incluyen este medio entre los contenidos de las *deepfakes* (S. Bray, Sh. Johnson y B. Kleinberg (2023), «Testing human ability to detect “deepfake” images of human faces». *Journal of Cybersecurity*, 2023, 1-18).

¹⁰ Esta definición está inspirada en Bray *et al.* (2023) que entienden que el texto entra dentro de los medios empleados por las *deepfakes*. «Lo importante es señalar que el contenido de una *deepfake* podría teóricamente ser cualquier cosa dentro de una determinada modalidad (video, imagen, audio o texto)». Véase *ibid.*, pág. 2.

¹¹ Por ejemplo, estas dos páginas web, que generan contenidos «*deepfakes*», ilustran adecuadamente el avance tan importante que los sistemas de IA generativa han alcanzado para producir *deepfakes* de altísima calidad:

- <https://is.gd/WdLDIY>.

- <https://is.gd/8cZaXk>.

¹² Le he pedido a ChatGPT que recree un discurso (falso, por supuesto) de Donald Trump apoyando el proyecto 2025 y este es el resultado: <https://is.gd/o9pYlt>.

de las *deepfakes* en relación con la democracia. El «mecanismo» por el cual las *deepfakes* producen un daño a nuestras democracias es más sutil de lo que pudiera parecer a primera vista. Examinémoslo de manera analítica. El punto de partida es un sistema de IA generativa que crea una *deepfake*, por ejemplo, un video de Kamala Harris diciendo que aceptó sobornos por parte de una importante compañía china de *software*. Ese video se planta en internet y en redes sociales. A partir de ahí, los usuarios de internet y de redes sociales lo visualizan y además lo diseminan por ambos. Los usuarios de internet y de redes sociales pueden visualizar y diseminar ese video fundamentalmente por dos razones:

- 1) porque desconozcan que el video es falso;
- 2) porque aun conociendo que el video es falso, quieren diseminarlo por razones de partidismo político.

Cuando decimos que el video tiene un «impacto negativo en la democracia» lo que queremos decir es, por tanto, que produce un impacto en aquellos que de forma honesta creerían en la veracidad de los contenidos que figuran en la red. Por definición, el video no produce ningún impacto negativo en aquellos que tienen una clara adscripción partidista y que, a sabiendas de que el video es falso, lo diseminan¹³. Aquellos usuarios que sí que creen en la veracidad de los contenidos que circulan por internet y por las redes sociales son, por tanto, *el objeto a proteger* en cualquier regulación en materia de *deepfakes*.

A su vez, dichos usuarios, proclives a creer en la veracidad de los contenidos que circulan por internet, pueden dividirse en dos grupos: aquellos que buscan información destinada a confirmar sus visiones políticas y aquellos que simplemente buscan información. Los primeros verán sus sesgos políticos confirmados por el video. A los segundos el video les producirá no tanto la confirmación de un sesgo político previo, sino más bien un sentimiento de desazón en relación con la política. Ambos grupos coincidirán, sin embargo, en el hecho de haber sido engañados y manipulados por el contenido del video. A pesar de ello, el tratamiento regulatorio de aquellos proclives al sesgo de confirmación es diferente del de aquellos que no están aquejados del mismo. En cualquier caso, el daño que se produce para la democracia, el coste fundamental en esta materia, es en términos de confianza¹⁴. Falta de confianza, en primer lugar, y en nuestro ejemplo, en un

¹³ Leonie Huddy está examinando el efecto del partidismo político versus otras explicaciones de tipo psicológico, como la capacidad de los usuarios de internet y redes sociales para detectar la credibilidad y exactitud de las noticias falsas y de las *deepfakes*. Véase L. Huddy (2024), «Sharing fake news: partisanship and accuracy as competing motives». Paper presentado en el Juan March Institute-UC3M permanent seminar, 24 de Noviembre 2023.

¹⁴ He explorado la cuestión de la relación entre confianza, derecho y democracia en, por ejemplo, A. Estella (2020), «Confianza institucional en América Latina: un análisis comparado». Fundación Carolina, Documento de Trabajo, 34/2020.

político importante, Kamala Harris. Falta de confianza, para continuar, en el resto de los políticos, puesto que está demostrado que nuestro modo de pensar funciona a través de asociaciones y estereotipos¹⁵: Kamala Harris es corrupta-Kamala Harris es un político-todos los políticos son corruptos. Finalmente, se produce una falta de confianza en la política como institución: toda la política, no ya solo los políticos, está podrida. Y el último paso es la falta de confianza en la democracia: puesto que la política es consustancial a la democracia, y la política está podrida, entonces la democracia también lo está.

El problema no solamente se plantea en la fase «ex ante», es decir, en la fase de creación y diseminación del video, sino en la fase «ex post», es decir, en la fase en la que se detecta que el video es una *deepfake*. En principio, la actividad orientada a detectar aquellos contenidos que son falsos es positiva, como no podría ser de otra manera. Pero como Chesney y Citron¹⁶ han argumentado de manera persuasiva, el descubrimiento de la falsedad puede generar un efecto que estos autores denominan «the liar's dividend»: puesto que sabemos que hay videos circulando por internet y por las redes sociales que son *deepfakes*, entonces empezamos a desconfiar de todos los contenidos que circulan por internet y por las redes sociales. Al final, el «dividendo» de aquel que genera y propaga la *deepfake* es todavía mayor, puesto que si lo que se pretende en última instancia es generar desconfianza en los políticos, la política, y finalmente, la democracia, una manera eficaz de hacerlo es consiguiendo que la gente piense que gran parte de la información que circula por internet y por las redes es falsa. En definitiva, el descubrimiento y la denuncia de *deepfakes* tiene efectos secundarios que pueden ser del todo indeseados.

Todos los análisis que se realizan en materia de *deepfakes*, y que intentan señalar el daño que la propagación de este tipo de contenidos tiene para la democracia, parte de una estructura similar a la que hemos planteado hace un momento. Es decir, se plantea un caso concreto, como hemos hecho nosotros con el ejemplo ficticio del video de Kamala Harris, y a partir de ahí se analiza el daño que ello puede producir a la democracia. Sin embargo, este tipo de análisis no deja de ser muy teórico: no sabemos cuál es el impacto *real* de las *deepfakes* en nuestra democracia. Salvo error u omisión por mi parte, nadie ha «medido» cuál podría ser ese impacto, ni nadie ha establecido una medida objetiva que nos permitiera saber cuál podría ser el mismo. Por ejemplo, no sabemos cuál es el número de *deepfakes* para hacer el mal a la democracia que circulan por las redes sociales, ni cómo podemos estimar el daño real que esas *deepfakes* que están circulando por internet y redes sociales está haciendo a la democracia. La discusión se mueve, por el momento al menos, en un nivel simplemente teórico. Parte de la agenda de investigación que se debería desarrollar en el futuro, debería estar

¹⁵ D. Kahneman (2024), *Thinking Fast and Slow*, Penguin Books .

¹⁶ B. Chesney y D. Citron (2019), «Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security». *California Law Review*, 107, 6, págs. 1753-1820.

centrada en «medir» en términos cuantitativos este fenómeno, el fenómeno de la erosión de la democracia por parte de las *deepfakes*¹⁷.

Por concluir con este apartado: el mecanismo a través del que las *deepfakes* operan es más sutil de lo que podamos pensar a primera vista. En realidad, las *deepfakes* erosionan no tanto la democracia, sino un aspecto vital en la democracia como es la confianza en los políticos, en la política y, finalmente, en la propia democracia. El objeto a proteger en este ámbito es aquellos usuarios que piensan de manera honesta que una *deepfake* incorpora contenidos veraces y, por tanto, la difunden. Esta categoría incluye aquellos que buscan confirmar sus sesgos políticos y aquellos que simplemente buscan la verdad e información creíble. Sin embargo, es evidente que estas dos categorías de personas deberían recibir un tratamiento diverso. Dejamos fuera de nuestro objeto de protección a aquellas personas que por razones de partidismo político diseminan las *deepfakes* a sabiendas, aunque sin duda alguna esta tercera categoría tiene que ser tenida en cuenta en cualquier regulación en materia de *deepfakes*. Adicionalmente, no hay que olvidar, este terreno, el llamado «liar's dividend», bajo mi punto de vista el riesgo más importante en este ámbito. Finalmente, aunque es verdad que desconocemos la medida del problema al que nos estamos enfrentado, es evidente que aquellos expertos que se han adentrado en la cuestión de la medición de las *deepfakes* y

¹⁷ Ilustro este punto con dos ejemplos. Ejemplo 1: en «The economics of *deepfakes*», un trabajo publicado en 2023, el autor señala que «al principio de 2023, el número de *deepfakes* online se estimó en el rango de millones». N. Kshetri (2023), «The Economics of *Deepfakes*». *IEEE COMPUTER SOCIETY*. Sin embargo, la estimación que hace este autor está basada en un artículo publicado en el *Wall Street Journal* el 25 de Febrero de 2023, titulado «The Deepfake Dangers Ahead», en el que se decía simplemente lo siguiente, sin proporcionar más datos ni citar la fuente: «en 2018, de acuerdo con un estudio [que no citan] menos de 10.000 *deepfakes* se detectaron online; en 2023, el número de *deepfakes* online era de millones» (D. Byman, Ch. Meserole y V.S. Subrahmanian (2023), «The Deepfake Dangers Ahead». *The Wall Street Journal* (23 de febrero). Ejemplo 2: en un estudio publicado en 2018, «Measuring the reach of «fake news and online disinformation in Europe» los autores, del Instituto Reuters de la Universidad de Oxford, señalan que «en la mayor parte de los casos analizados en este estudio, tanto en Francia como en Italia, las webs de «fake news» no generaron tantas consultas como las páginas webs de marcas de noticias establecidas, como Le Monde o la Repubblica». Al mismo tiempo, estos investigadores advierten de que «carecemos de incluso la información más básica sobre la dimensión del problema en materia de fake news en todos los países» (R. Fletcher, A. Cornia, L. Graves y R. Nielsen (2018), «Measuring the reach of fake news and online disinformation in Europe», Factsheet. February, Reuters Institute. University of Oxford, pág. 1). Es importante señalar que, en este último estudio, estamos hablando de «fake news», y no de *deepfakes*, aunque sin duda alguna ambas cuestiones están co-relacionadas. Sorprende con carácter general que la cuestión de la medición del número de *deepfakes* que circulan *online* no haya recibido una atención más profunda, lo que hace difícil establecer una medida del impacto real de las *deepfakes* en la sociedad y en particular en nuestras democracias.

de su impacto en la sociedad nos están advirtiendo de que el problema tiene una magnitud, probablemente, muy importante.

III. LA REGULACIÓN DE LOS SISTEMAS DE IA QUE NO ESTÁN PROHIBIDOS Y QUE NO CONSTITUYEN RIESGOS ALTOS; EN PARTICULAR, LA REGULACIÓN DE LAS DEEPFAKES

El reglamento de IA aborda la regulación de los sistemas de IA¹⁸ desde la perspectiva (errónea, bajo mi punto de vista, como veremos a continuación) de los riesgos que los mismos son capaces de producir. En este sentido, establece una trilogía¹⁹: riesgos inaceptables, y por tanto, prohibidos²⁰; riesgos altos, a los que somete a una serie de requisitos y obligaciones²¹; y «determinados sistemas de IA» (es decir, el resto) a los que somete a obligaciones de transparencia²².

Vamos a centrarnos en esta sección en la regulación de los demás sistemas de IA (de aquellos que no constituyen riesgos inaceptables ni riesgos altos), dentro de la cual están reguladas las *deepfakes*. El reglamento dedica el capítulo IV, y dentro de él, el art. 50, al resto de sistemas de IA, es decir, a aquellos que presentan riesgos aceptables, que el reglamento denomina como «determinados sistemas de IA». El título de este capítulo («obligaciones de transparencia de los proveedores y responsables de determinados sistemas de IA») indica claramente cuál es la regulación prevista en este ámbito: se imponen obligaciones fundamentalmente de información a los proveedores de sistemas de IA, y también a los responsables del

¹⁸ Dejo a un lado la cuestión de los modelos de IA de uso general, que el reglamento también regula (aunque se centra más profusamente en los sistemas de IA). Los modelos de IA de uso general son definidos por el reglamento en su art. 3.63 como «un modelo de IA, también uno entrenado con un gran volumen de datos utilizando supervisión a gran escala, que presenta un grado considerable de generalidad y es capaz de desarrollar de manera competente una gran variedad de tareas distintas, independientemente de la manera en que el modelo se introduzca en el mercado, y que puede integrarse en diversos sistemas o aplicaciones posteriores, excepto los modelos de IA que se utilizan para actividades de investigación, desarrollo o creación de prototipos antes de su introducción en el mercado». Un modelo de IA puede ser un componente de un sistema de IA, pero no al revés. Por ejemplo, GPT es el modelo que se emplea en ChatGPT, que es, este, un sistema de IA. Por emplear una metáfora, un modelo sería como el motor de un vehículo, mientras que el vehículo sería el sistema de IA.

¹⁹ Véase con carácter general A. Huergo (2023), «Inteligencia artificial: una aproximación jurídica no catastrofista», *Revista Española de Control Externo*, XXV, 74-75 (mayo-septiembre) | págs. 110-129, y E. Gamero (dir.) y F. L. Pérez Guerrero (2023), *Inteligencia Artificial y Sector Público*. Valencia: Tirant lo Blanch.

²⁰ Capítulo II, art. 5.

²¹ Capítulo III, arts. 6 al 27.

²² Capítulo IV, art. 50.

despliegue de determinados sistemas de IA. El reglamento define en su art. 3.4 lo que hay que entender por «responsable del despliegue»: se trata de una «persona física o jurídica, o autoridad pública, órgano u organismo que utilice un sistema de IA bajo su propia autoridad, salvo cuando su uso se enmarque en una actividad personal de carácter no profesional». Por su parte, el reglamento define al proveedor, en su art. 3.3º, como «una persona física o jurídica, autoridad pública, órgano u organismo que desarrolle un sistema de IA o modelo de IA de uso general y lo introduzca en el mercado o para el que se desarrolle un sistema de IA o un modelo de IA de uso general y lo introduzca en el mercado o ponga en servicio el sistema de IA con su propio nombre o marca, previo pago o gratuitamente».

La estructura del art. 50 del reglamento es la siguiente: el art. 50 establece 5 supuestos; además, hay que señalar que en realidad el art. 50 establece un mínimo que todos los sistemas de IA deben cumplir²³. En efecto, como señala el art. 50. 6 «los apartados 1 a 4 [del art. 50] no afectarán a los requisitos y obligaciones establecidos en el Capítulo III y se entenderán sin perjuicio de otras obligaciones de transparencia establecidas en el derecho nacional o de la Unión para los responsables del despliegue de sistemas de IA». Lógicamente, la referencia se hace a los sistemas de IA de alto riesgo, puesto que por definición, los sistemas de IA que constituyen riesgos inaceptables, al estar prohibidos por el reglamento, no se les aplica obligación de transparencia alguna.

Como digo, el reglamento de IA establece 5 supuestos, que son los siguientes:

- Sistemas de IA destinados a interactuar directamente con personas físicas.
- Sistemas de IA de uso general que generen contenido sintético de audio, imagen, video, texto.
- Sistemas de reconocimiento de emociones o de categorización biométrica.
- Sistemas de IA que generen o manipulen imágenes o contenidos de audio o video que constituyan una *ultrasuplantación*
- Sistemas de IA que generen o manipulen texto que se publique con el fin de informar al público sobre asuntos de interés público²⁴.

Vamos a examinar con detalle cada uno de estos supuestos. Pero antes de entrar en ello, es importante dilucidar la cuestión de si estas obligaciones son

²³ Así lo entiende también Th. Gils (2024, pág. 31). «A Detailed Analysis of Article 50 of the EU's AI Act», Próxima publicación en C. N. Pehlivan, N. Forgó y P. Valcke (eds.), *The EU Artificial Intelligence (AI) Act: A Commentary* (Kluwer Law International), pág. 31.

²⁴ Sin embargo, este quinto caso se incluye en el párrafo 4º del art. 50, dedicado a la regulación, específicamente, de las *deepfakes*. Esta sistemática es confusa, y si acaso, ilustra perfectamente la necesidad de ampliar la definición de las *deepfakes* de tal manera que se englobe la generación y manipulación de texto.

cumulativas o no. Desde mi punto de vista, y siguiendo en este punto a Gils²⁵, dichas obligaciones son efectivamente cumulativas, puesto que el reglamento no indica nada *a contrario*; lo que implica que, por ejemplo, una *deepfake* tendrá que cumplir con las obligaciones establecidas en el art. 50.4, propias de la regulación de las *deepfakes*, además de con las obligaciones establecidas en el resto de los párrafos de dicho artículo. Hay una segunda razón por la que se impone la acumulación de obligaciones, y es que, como veremos más adelante, es difícil distinguir nítidamente estos 5 casos entre sí. Por tanto, ante dicha dificultad, el criterio más prudente parece el de la acumulación.

1. SISTEMAS DE IA DESTINADOS A INTERACTUAR DIRECTAMENTE CON PERSONAS FÍSICAS

Pensemos, por ejemplo, en asistentes de voz (como, por ejemplo, Alexa cuando se terminen de integrar sistemas de IA en este asistente de voz) o en los llamados «robotic companion pets» (robots-mascotas de compañía). Todos estos sistemas de IA están «destinados», dice el reglamento, a interactuar directamente con personas. Pues bien, el reglamento obliga al proveedor a informar a las personas que los empleen de que están interactuando con sistemas de IA.

Lo primero que hay que tener en cuenta a este respecto es que el proveedor del sistema de IA, aunque sea el introductor del mismo en el mercado, puede no ser el productor del producto que sea adquirido por el consumidor final. A pesar de ello, el proveedor será el responsable de que el productor efectivamente informe de que se está empleando un producto que tiene integrado un sistema de IA. Por ejemplo, el proveedor de un sistema de IA que luego sea empleado por Amazon para fabricar Alexa es el responsable de asegurarse de que Amazon efectivamente informe al consumidor final de que Alexa está integrado por un sistema de IA. Ello genera una disrupción jurídica que seguramente será objeto de una profusa litigiosidad en el futuro. Por otro lado, el art. 50.1 no indica en qué debe consistir esa obligación de información, exactamente. En este sentido, el preámbulo del reglamento no da indicaciones adicionales a este respecto, a diferencia de lo que ocurre con otros casos, como veremos más adelante. Gils²⁶ sobre la base de un análisis sistemático del párrafo 132 del preámbulo del reglamento (en el que este caso, sistemas de IA que interactúan directamente con humanos, se incluye) propone que esta obligación de información se concrete en una obligación de «notificar» a la persona que se está interactuando con un sistema de IA. Independientemente de cómo podamos valorar esta hermenéutica, lo cierto es que tampoco queda muy claro a qué se referiría, específicamente, esta obligación de notificación, si es que esta fuera la interpretación

²⁵ Th. Gils (2024, pág. 8).

²⁶ Th. Gils (2024, pág. 89).

escogida. Lo que es evidente es que se debe «comunicar», como dice el párrafo 132, a las personas que interactúen directamente con un sistema de IA, que este es el caso.

Además, hay que señalar que el reglamento establece una excepción para este supuesto, y es que «resulte evidente desde el punto de vista de una persona razonablemente informada, *atenta y perspicaz*, teniendo en cuenta las circunstancias y el contexto de utilización» que está interactuando con un sistema de IA. Esta excepción es objeto de evidentes críticas, puesto que parece difícil establecer un criterio objetivo a este respecto. Se deja, por tanto, que sea el aplicador del derecho (administración, y en su caso, tribunales de justicia)²⁷ el que module la extensión de dicha excepción (por ejemplo, lo que puede ser evidente para un profesor de derecho experto en Inteligencia Artificial puede no serlo para un niño de 10 años, o para una persona con dificultades). Se abre una clara ventana de potencial litigiosidad, que se hubiera podido cerrar, simplemente, eliminando esta surrealista excepción. Finalmente, hay que señalar que la obligación no se aplica a aquellos casos de sistemas de IA autorizados por ley para el llamado «law enforcement», es decir, para detectar, prevenir, investigar o enjuiciar delitos. Sin embargo, a su vez, esta excepción conoce una nueva excepción, puesto que el reglamento indica que en los casos en los que estos sistemas estén a disposición del público para denunciar un delito penal, será aplicable la obligación de informar de que se está interactuando con un sistema de IA (imaginemos por ejemplo el caso de una web de acceso público que emplee un chatbot para la prevención de delitos). En definitiva, el art. 50.1 introduce un complejo sistema de excepciones que planteará problemas de aplicabilidad de esta disposición, lo que se hubiera podido evitar simplemente con el establecimiento de una obligación generalizada de información, que podría tener alguna excepción en el ámbito penal.

2. SISTEMAS DE IA DE USO GENERAL QUE GENEREN CONTENIDO SINTÉTICO DE AUDIO, IMAGEN, VIDEO, TEXTO

El primer problema que plantea este segundo supuesto es el de qué debemos de entender por «contenido sintético». En efecto, el reglamento no define legalmente este concepto. De acuerdo con la AEPD (Agencia Española de Protección de Datos), los datos sintéticos «son generados artificialmente, a diferencia de los datos reales que se recopilan en la realidad. Un conjunto de datos no reales, que se etiquetará como datos sintéticos, debe preservar las características y propie-

²⁷ S. de la Sierra (2020), «Inteligencia artificial y justicia administrativa: una aproximación desde la teoría del control de la Administración Pública», *Revista General de Derecho Administrativo*, 53.

dades de los datos reales para un caso de uso específico»²⁸. Los datos sintéticos se emplean habitualmente por los sistemas de IA generativa²⁹. En este sentido, Thomson³⁰ estima que alrededor del 70% de los datos empleados para entrenar a GPT-5 serían datos sintéticos. Esto es probablemente lo que le lleva a Gils³¹ a señalar que este caso se está refiriendo en realidad a lo que «se llama comúnmente IA generativa». Hay que señalar que en lo que constituye una verdadera laguna del reglamento (un agujero negro, más bien habría que decir), el reglamento de IA no aborda directamente la regulación de la IA generativa, lo que es duramente criticado por Wachter³². En cualquier caso, y aun siendo conscientes de que probablemente no sea del todo exacto equiparar contenidos sintéticos con IA generativa, lo cierto es que ambos están muy relacionados, por lo que seguiremos en lo que sigue la posición de Gils, y entenderemos que este supuesto se refiere fundamentalmente a la IA generativa. Por ejemplo, ChatGPT entraría dentro de esta categoría, en lo que se refiere a la generación de texto.

Pues bien, en relación con este supuesto, el art. 50.2 del reglamento establece una obligación de «marcado». En efecto, los resultados «de salida» del sistema de IA deben de estar marcados en un formato «legible por máquina» de tal manera que sea posible detectar que han sido «generados o manipulados» de manera artificial. Gils³³ llama la atención sobre las contradicciones del reglamento en este sentido. En efecto, mientras la primera frase del art. 50.2 se refiere a los proveedores que «generen contenido sintético», la tercera frase se refiere, como hemos visto, tanto a la generación como a la manipulación. Por tanto, hay que entender que la obligación de marcado se impone al proveedor que genere y también al que manipule contenido sintético. Es decir: los contenidos generados y manipulados resultado del uso de sistemas generativos de IA que empleen datos sintéticos deben de estar marcados por igual.

En lo que se refiere, específicamente, a los sistemas de marcado, hay que traer a colación el párrafo 133 del preámbulo del reglamento. En efecto, el párrafo 133 cita, específicamente a las «marcas de agua, identificación de metadatos, métodos criptográficos para demostrar la procedencia y la autenticidad del contenido,

²⁸ AEPD, «Datos sintéticos y protección de datos». Consultar en: <https://is.gd/LVupWg>. Último acceso el 20 de octubre de 2024.

²⁹ Como explica Joshi, en «Synthetic AI vs. Generative AI: which one to use to strengthen data engineering in machine learning». Consultar en: <https://is.gd/x6a6x8>. Último acceso: 20 de octubre de 2024.

³⁰ Alan Thompson (2024), *What's in GPT-5? A Comprehensive Analysis of Datasets Likely Used to Train GPT-5*, LifeArchitect.ai, agosto 2024.

³¹ Th. Gils (2024, pág. 11).

³² S. Wachter (2024, pág. 694), «Limitations and Loopholes in the EU Act and AI Liability Directives: What this means for the EU, the USA, and beyond». *Yale Journal of Law and Technology*, vol. 26, Issue 3.

³³ Th. Gils (2024, pág. 12).

métodos de registro, impresiones dactilares, u otras técnicas, según proceda». Gils nos informa de que «el objetivo primordial de estas técnicas no es precisamente el de permitir la distinción entre contenido artificial o humano [...] sino más bien la autenticación del contenido y su proveniencia». Por ejemplo, en este sentido, Madiega³⁴ señala que «en la práctica, la marca de agua crea una firma única identificable que es invisible para los humanos pero que es detectable algorítmicamente y que puede permitir la trazabilidad del modelo de IA». Por otro lado, Gils se pregunta si esta lista es exhaustiva o no. Entiendo que no hay que interpretarla como un *numerus clausus*, puesto que la frase final del párrafo 133 antes transcrito deja la puerta abierta a otras técnicas de marcado al emplear la expresión «u otras técnicas, según proceda».

3. SISTEMAS DE RECONOCIMIENTO DE EMOCIONES O DE CATEGORIZACIÓN BIOMÉTRICA

En relación con este supuesto, hay que empezar señalando que el reglamento define los sistemas de reconocimiento de emociones y de categorización biométrica en sus arts. 2.39 y 2.40, respectivamente. Lo primero que llama la atención en relación con ambos subcasos es la deficiente técnica legislativa empleada por el legislador comunitario. Efectivamente, en ambos casos estamos hablando de sistemas de IA, a pesar de que la expresión empleada por el art. 50.3 pueda llevarnos a pensar que estamos hablando de sistemas que no sean solamente de IA. Así, el art. 2.39 dice que los «sistemas de reconocimiento de emociones son sistemas de IA destinados a distinguir o inferir las emociones o las intenciones de las personas físicas a partir de sus datos biométricos»; mientras que el art. 2.40 define a los sistemas de categorización biométrica como «sistemas de IA destinados a incluir a la persona física en categorías específicas en función de sus datos biométricos, a menos que sea accesorio a otro servicio comercial y estrictamente necesario por razones técnicas objetivas». Estamos hablando, pues, en ambos casos, de sistemas de IA.

La obligación fundamental que establece el art. 50.3 es la de informar del funcionamiento del sistema a las personas físicas expuestas a él, por un lado, y por otro, a tratar sus datos personales de conformidad con los reglamentos 2016/679 y 2018/1725 y la Directiva 2016/680. No entro en las consideraciones de protección de datos en este trabajo. Por lo que se refiere a la primera de las obligaciones, una de las primeras diferencias con respecto a los otros dos casos que hemos examinado hasta ahora es que el sujeto al que se impone dicha obligación no es el proveedor del sistema de IA sino aquellas personas que desplieguen el mismo (es decir, de acuerdo con la definición del reglamento, aquellas personas que utilicen el sistema de IA). No queda claro por qué en estos casos se impone la obligación solamente a los «desple-

³⁴ T. Madiega (2023, pág. 2): «Generative AI and watermarking», European Parliament Briefing.

gadores» del sistema de IA y no a los proveedores, mientras que en los demás casos se imponen a los proveedores, lo que es más coherente con la sistemática general del reglamento, que como hemos dicho antes, sitúa con carácter general el centro de imputación de la responsabilidad del cumplimiento de las obligaciones que establece el reglamento en la persona del proveedor. Gils³⁵ comenta en relación con esta obligación que no queda claro si de lo que hay que informar a la persona expuesta a estos sistemas es de que está siendo expuesta (interpretación restrictiva) o de que está siendo expuesta a los mismos y, adicionalmente de cómo funcionan dichos sistemas (interpretación amplia). El autor parece que se decanta por la segunda interpretación, amplia, teniendo en cuenta el carácter invasivo de ambos sistemas de IA. Sin embargo, no parece que sea esa la lógica del sistema establecido por el art. 50 del reglamento, el cual establece, como estamos viendo, meras obligaciones de información de que se está ante sistemas de IA en distintos ámbitos, y no de su forma de funcionar. Sin duda alguna, las autoridades de gobernanza administrativa del reglamento de IA tendrán que aclarar esta cuestión, en particular la Oficina de IA de la Comisión³⁶, como el propio Gils señala. Por otro lado, el párrafo 132 del preámbulo del reglamento impone una obligación de «notificación» a las personas expuestas a sistemas de IA de reconocimiento de emociones a través de datos biométricos, por lo que parece que es ésta la obligación específica de información que se establece para los sistemas de IA de reconocimiento de emociones o de categorización biométrica con carácter general. Finalmente, hay que señalar que autores como Wachter³⁷ critican duramente que los sistemas de IA de reconocimiento de emociones y de categorización biométrica queden fuera de los sistemas de IA de alto riesgo.

4. LAS DEEPFAKES, EN PARTICULAR: SISTEMAS DE IA QUE GENEREN O MANIPULEN IMÁGENES O CONTENIDOS DE AUDIO O VIDEO QUE CONSTITUYAN UNA ULTRASUPLANTACIÓN

Llegamos de esta manera a la regulación, específicamente, de las *deepfakes*, el objeto primordial de este artículo. El art. 50.4 establece literalmente que «los responsables del despliegue de un sistema de IA que genere o manipule imágenes o contenidos de audio o video que constituyan una ultrasuplantación harán público que estos contenidos o imágenes han sido generados o manipulados de manera artificial».

³⁵ Th. Gils (2024, pág. 20).

³⁶ El art. 3.47 del reglamento de IA define a la Oficina de IA como «la función de la Comisión consistente en contribuir a la implantación, el seguimiento y la supervisión de los sistemas de IA y modelos de IA de uso general, y a la gobernanza de la IA prevista por la Decisión de la Comisión de 24 de Enero de 2024».

³⁷ S. Wachter (2024, pág. 680).

Lo primero que tenemos que destacar es que el reglamento hace responsable de la obligación de informar al responsable del despliegue del sistema de IA que genere la *deepfake*, y no al proveedor. Es decir, se hace responsable a aquellas personas que utilicen el sistema de IA generador de la *deepfake*. El segundo punto que debemos destacar es que se obliga a dichos sujetos a que «hagan público» que esos contenidos han sido generados o manipulados de manera artificial. Sin embargo, el reglamento no señala de manera específica cómo se debe implementar dicha obligación, es decir, cómo, exactamente, se debe hacer público que una determinada *deepfake* ha sido generada o manipulada por un sistema de IA. En este sentido, hay que traer a colación el párrafo 134 del preámbulo, puesto que el mismo da alguna que otra indicación al respecto. Concretamente, dicho párrafo nos dice que el responsable del despliegue debe «etiquetar» los resultados de salida generados por IA «en consecuencia» e indicar su origen artificial. El párrafo 134 añade: «también conviene prever una obligación de divulgación similar en relación con el texto generado o manipulado por una IA en la medida en que se publique con el fin de informar al público sobre asuntos de interés público». Por tanto: la obligación de «hacer público» que un determinado contenido es una *deepfake* se concreta en una obligación de etiquetado, obligación que se extiende a los contenidos generados o manipulados de texto, como veremos más adelante.

Adicionalmente, la primera frase de dicho párrafo establece que la anterior obligación se establece «además de las soluciones técnicas utilizadas por los proveedores del sistema de IA [...]». Gils³⁸ interpreta en un sentido acumulativo dicha frase, es decir: concretamente, a la obligación de etiquetado, le sobrevendría, por ejemplo, la obligación de marcado que hemos visto en relación con el art. 50.2 (contenidos sintéticos).

En relación con esta obligación de etiquetado, Gils señala que, a pesar de que el reglamento no dice cómo, exactamente, hay que etiquetar que una *deepfake* es tal, es evidente que la obligación de etiquetado tiene que ser perceptible a los ojos humanos. Más allá de ello, el reglamento guarda silencio en relación con la cuestión de qué debería decir la etiqueta de la *deepfake* y cómo debería, exactamente, presentarse dicha etiqueta. Es decir, esta cuestión plantea a su vez dos subcuestiones, una formal y otra sustantiva. La formal es qué forma debería adoptar la etiqueta, y la sustantiva es qué contenido(s) debería tener concretamente la etiqueta de *deepfakes*. En este sentido, Van Huijstee *et al*³⁹ señalan que este aspecto del reglamento debería ser desarrollado «a través de orientaciones» (*guidelines*). Parece que estos autores apuntan hacia la solución de someter la

³⁸ Th. Gils, (2024, pág. 23).

³⁹ Van Huijstee (2021, pág. 87) en «European Parliament (Panel for the Future of Science and Technology) (2021), «Tackling *Deepfakes* in European Policy». Study. July 2021.

clarificación de esta obligación a procesos de estandarización⁴⁰. Sin embargo, me parece que la cuestión es tan importante que, independientemente de que se recabe la opinión de la industria y los interesados en materia de *deepfakes*, debería ser la Comisión Europea la que, a través de actos legislativos delegados, o actos de ejecución, regulara la materia. Es decir, la concreción de esta obligación no debería dejarse al ámbito del *soft-law*.

Adicionalmente, Gils⁴¹ indica que el Reglamento de Servicios Digitales⁴², concretamente el art. 35.1 k) de dicho reglamento, entra en juego aquí. En efecto, dicho artículo establece que los proveedores de plataformas en línea de muy gran tamaño y de motores de búsqueda en línea de muy gran tamaño deben «garantizar que un elemento de información [...] que se asemeja notablemente a personas, objetos, lugares (etc.) y que puede inducir erróneamente a una persona a pensar que es auténtico o verídico, se distinga mediante indicaciones destacadas cuando se presente en sus interfaces en línea, y además, [proporcione] una funcionalidad fácil de utilizar que permita a los destinatarios del servicio señalar dicha información». Por ejemplo, la plataforma YouTube debería no solamente garantizar que se indique cuándo un contenido es falso, sino que además debe de establecer una herramienta que permita que el que suba un video a esta plataforma, pueda señalar de manera relativamente sencilla que ese contenido es falso. El reglamento se aplica con carácter general a los fakes, es decir, a los contenidos falsos, y no estrictamente a las *deepfakes*, pero es evidente que el las *deepfakes* son una especie del género, más amplio, fakes. Pues bien, el reglamento de IA no aclara si con esta advertencia sería suficiente, o si habría que añadir a esta advertencia un etiquetado específico en relación con la *deepfakes*. Puesto que nos falta por saber cómo se va a desarrollar esta obligación de etiquetado en el futuro no podemos responder en estos momentos a esta pregunta. Es evidente que lo importante, a fin de cuentas, es que la persona quede protegida y conozca o pueda conocer de manera clara y sin demasiados esfuerzos que un determinado contenido es una *deepfake*, al mismo tiempo que este tipo de obligaciones no adquieren un carácter redundante y por tanto excesivamente costoso para el proveedor de sistemas de IA capaces de generar *deepfakes* y el responsable del despliegue de *deepfakes*.

El reglamento establece dos excepciones a esta obligación de etiquetado. La primera es para los casos de «law enforcement», es decir: no hay que etiquetar cuando «la ley autorice su uso [el uso de *deepfakes*] para detectar, prevenir, inves-

⁴⁰ Véase con carácter general H. Schepel (2005), *The Constitution of Private Governance. Product Standards in the Regulation of Integrating Markets* Hart Publishing, y B. de Witte (2023), «Soft law in European public law» en *Research Handbook on Soft Law*. Edward Elgar Publishing.

⁴¹ Th. Gils (2024, pág. 23).

⁴² Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo de 19 de octubre de 2022 relativo a un mercado único de servicios digitales y por el que se modifica la Directiva 2000/31/CE (Reglamento de Servicios Digitales)

tigar o enjuiciar delitos». La segunda excepción es la referida a «obras o programas manifiestamente creativos, satíricos, artísticos, de ficción o análogos»; en este caso, las obligaciones de transparencia establecidas en el art. 50.4 se modularán, de tal manera que la obligación de información no dificulte la exhibición o el disfrute de la obra. De nuevo, el reglamento no da ninguna indicación de cómo se debe informar en estos casos, de cuáles deben ser los contenidos de dicha información y la forma de proveerlos, por lo que estamos, de nuevo, en un ámbito que necesitará de desarrollo normativo por parte de la Comisión Europea. Las reflexiones que deslizaba más arriba, en el sentido de que debe ser la Comisión Europea, es decir, el gobierno y la Administración pública comunitaria, y no los organismos de estandarización, la que actúe en este ámbito, son aplicables aquí *mutatis mutandi*.

5. SISTEMAS DE IA QUE GENEREN O MANIPULEN TEXTO QUE SE PUBLIQUE CON EL FIN DE INFORMAR AL PÚBLICO SOBRE ASUNTOS DE INTERÉS PÚBLICO

Es interesante observar cómo el art. 50.4 regula, en su segundo párrafo, los contenidos de texto, generados o manipulados por sistemas de IA de manera artificial. Este aspecto refuerza la tesis que he defendido más arriba, en este trabajo, según la cual una definición jurídica correcta de las *deepfakes* debería englobar los contenidos de imagen, video, audio y también texto. En relación con contenidos de texto, la regulación es simétrica a la del resto de las *deepfakes*, analizadas más arriba. Es decir: se impone la obligación «similar», dice el reglamento, de informar por parte del responsable del despliegue; la obligación es, como digo, una obligación de «divulgar» que el texto se ha generado o manipulado de manera artificial; y dicha obligación es tan inconcreta como en el caso de las *deepfakes*, aunque por la simetría con la regulación de las ultrasuplantaciones podemos interpretar que se trata, también, de una obligación de etiquetado.

Parecería, por tanto, que la lógica del reglamento es que hay ultrasuplantación cuando hablamos de imagen, video y audio (por ejemplo, de una persona), pero no puede haber ultrasuplantación cuando lo que hablamos es de texto. Sin embargo, es evidente que también en este caso podemos hablar de ultrasuplantación: imaginemos que ChatGPT elabora un texto como si fuera un texto de Thomas Mann, sin serlo. No se ve claramente por qué no podemos denominar a esta modalidad como ultrasuplantación⁴³. El reglamento introduce un matiz

⁴³ He pedido a ChatGPT que genere un texto inspirado en el libro de Thomas Mann *La montaña mágica* y este es el resultado. ChatGPT aclara que «este fragmento es completamente generado por mí y no aparece en *La montaña mágica* de Thomas Mann. Lo escribí inspirado en su estilo, usando temas y el tono que son característicos de la obra, pero es

a este respecto y es que la finalidad de ese texto «falso» sea la de informar al público de asuntos de interés público, como si estuviera haciendo una división entre *deepfakes* «for good» y *deepfakes* «for bad» (que en el caso del texto «for bad», se regularían a través del art. 50.2, como contenidos sintéticos⁴⁴). Pero pensemos en que el texto es generado o manipulado sin que se empleen contenidos sintéticos; en ese caso, la *deepfake* textual quedaría fuera de la regulación del reglamento. En cualquier caso, lo importante sería proteger -y parece que el reglamento obvia este punto- al humano receptor del texto. Por otro lado, el reglamento establece una excepción en este caso, que es que «el contenido [es

una creación original, no una cita ni una reproducción de ningún pasaje del libro. Intenté capturar el espíritu contemplativo y existencial que se encuentra en el texto de Mann, pero sin copiar ningún contenido específico. ¡Así que no te preocupes! Este es solo un ejemplo ficticio, no una reproducción literal». El texto generado artificialmente es el siguiente.

Fragmento inspirado en *La montaña mágica* generado artificialmente por ChatGPT el 29 de octubre de 2024 a las 12:21 horas:

«Hans Castorp se encontraba una vez más en el balcón de su habitación, observando cómo la niebla descendía lenta y deliberadamente desde las cumbres, como un manto impasible y ancestral. Era un día húmedo, como tantos otros en aquel sanatorio, y el silencio era casi palpable, un eco sin fin de las soledades del espíritu. Al tiempo que la bruma borraba las líneas del paisaje, sentía cómo, poco a poco, se desdibujaban también las fronteras de su propia identidad.

Lejos de aquella altitud aislada, el mundo continuaba girando, gobernado por leyes y ritmos mundanos; pero en la alta soledad de Davos, el tiempo era una sustancia distinta, maleable y casi líquida. Castorp miró el reloj en su muñeca y, al ver la aguja apenas moverse, le pareció que el tiempo no pasaba: era él quien era absorbido por aquel flujo invisible, casi como una gota de agua que desaparece en el vasto océano de la eternidad.

“¿Qué significa estar aquí, viviendo tan próximo a la nada?”, se preguntaba en un murmullo, consciente de que las respuestas no estaban al alcance de su entendimiento, pero seducido por la insistente necesidad de hallar alguna señal en las profundidades de aquella niebla. Se sentía una pieza en una vasta y absurda maquinaria, un peón en el tablero de una partida cósmica cuyo sentido le era ajeno, aunque lo contenía.

La puerta de su habitación crujió suavemente, y la figura esbelta del doctor Behrens apareció, trayendo consigo el aroma medicinal de la autoridad y la serenidad del saber. “Señor Castorp”, dijo el doctor, con la voz grave y pausada de quien ha visto los límites de la condición humana, “quizá encuentre respuestas si deja de buscarlas con tanta impaciencia”.

⁴⁴ Esta sería la única opción que nos dejaría el reglamento, puesto que ya hemos visto que el art. 50.2 sí que se refiere a los contenidos de texto (eso sí, sintéticos). El reglamento somete a una obligación de marcado aquellos contenidos sintéticos de texto que hayan sido generados o manipulados de manera artificial. Aquellos que hubieran sido generados con una finalidad de información pública caerían bajo la órbita del art. 50.4. Como decimos, aquellos contenidos de texto que no fueran de carácter sintético generados no para una finalidad de información pública, sino para otras (por ejemplo, hacer el mal) no estarían regulados por el reglamento.

decir, el texto] generado por IA haya sido sometido a un proceso de revisión humana o de control editorial y cuando una persona física o jurídica tenga la responsabilidad editorial por la publicación del contenido». Es decir, el reglamento establece dos posibilidades: 1) que el texto haya sido revisado por un humano; 2) que el texto haya sido controlado a través de una labor editorial⁴⁵. De nuevo, parece más sencillo articular la excepción cuando se trate de textos falsos que estén bajo control, por ejemplo, de una editorial, pero es menos evidente cómo se articulará la primera excepción: ¿cómo se proporciona la prueba de que el contenido de texto ha sido generado o manipulado bajo el control de un humano?

Finalmente, hay que señalar un último aspecto que afecta a toda la estructura del art. 50. Como dice Gils⁴⁶, la manera correcta de entender la cascada de obligaciones que establece el art. 50 es de manera acumulativa, puesto que el art. 50 no dice nada que permita pensar que estas obligaciones se aplican de manera alternativa e independiente entre sí. Esto afecta de manera muy directa a la cuestión de las *deepfakes*; por tanto, la generación o manipulación de audio, video y voz (y texto, desde mi perspectiva) con el resultado de crear una *deepfake* puede estar sometido a todas estas obligaciones:

- Una obligación de comunicación (que puede concretarse en una obligación de notificación) si el sistema de IA está destinado a interactuar directamente con personas (que se impone al proveedor).
- Una obligación de marcado, si se han empleado contenidos sintéticos para generar la *deepfake* (que se impone al proveedor).
- Y finalmente, una obligación de etiquetado de la *deepfake*⁴⁷ (que se impone al responsable del despliegue del sistema de IA que genere la *deepfake*).

⁴⁵ El reglamento establece la consabida excepción relativa al «law enforcement»: cuando el texto se genere o manipule artificialmente por un sistema de IA y dicha generación o manipulación esté autorizada por ley para detectar, prevenir, investigar o enjuiciar delitos, la obligación de divulgación no se aplicará.

⁴⁶ Th. Gils (2024, pág. 8).

⁴⁷ No parece que los sistemas de reconocimiento de emociones y de reconocimiento biométrico puedan ser generadores de *deepfakes* cuando se emplean para su propio propósito, por lo que he excluido la mención de las obligaciones que corresponden a estos sistemas. Sin embargo, podíamos pensar en sistemas de reconocimiento de emociones que devolvieran al humano emociones suplantadas al producirse el reconocimiento, y sistemas de reconocimiento biométrico que devolvieran al humano que está siendo reconocido imágenes suplantadas del mismo. En ese caso, las obligaciones impuestas en esta área podrían entrar en juego también, junto con todas las demás que he mencionado antes.

De todas estas obligaciones, es evidente que la más importante de ellas es la del etiquetado.

IV. LA APLICACIÓN DEL ART. 50.4 DEL REGLAMENTO DE IA

¿Qué consecuencias jurídicas tendría la introducción en las redes de un *deepfake* con el objetivo de desestabilizar la democracia (pensemos en el video de Kamala Harris del principio de este artículo), desde un punto de vista jurídico? Lo primero que hay que examinar para responder a esta pregunta es la cuestión competencial. El reglamento establece una división de competencias en el ámbito de ejecución del mismo, entre las autoridades administrativas nacionales y las comunitarias. En principio, el reglamento atribuye competencia directamente a la Comisión Europea para la aplicación del reglamento en lo que se refiere a los modelos de IA que presentan riesgo sistémico, mientras que para el resto de los casos (es decir, sistemas de IA que presenten diversos riesgos) se atribuye competencia a las autoridades administrativas nacionales. Sin embargo, hay que dejar bien claro y desde el principio que el reglamento solamente contempla la generación de *deepfakes* por parte de sistemas de IA, y no de modelos de IA. En consecuencia, la competencia para actuar recaerá fundamentalmente en los Estados miembros. Sin embargo, es evidente que se pueden generar y/o manipular *deepfakes* por parte de modelos de IA, pero, como decimos, el reglamento, simplemente, no contempla dicha posibilidad, en lo que constituye una clara laguna jurídica⁴⁸.

Por lo tanto, si se generara y/o manipulara una *deepfake* por parte de un sistema de IA, el reglamento establece, en sus arts. 79 a 83, el procedimiento que hay que seguir para corregir la posible vulneración del art. 50 (lo que atañe fundamentalmente, por lo que decimos, a las autoridades nacionales). En este sentido, el reglamento establece, en su art. 79.6, lo siguiente:

[...] la notificación a la que se refiere el apartado 5 incluirá todos los detalles disponibles [...]. En concreto, las autoridades de vigilancia del mercado indicarán si la no conformidad se debe a uno o varios de los motivos siguientes [...]:

⁴⁸ Este será, por otro lado, el caso habitual, puesto que los modelos de IA «alimentan» a los sistemas de IA. Por tanto, lo normal será que los modelos de IA formen parte de los sistemas de IA y que sean los sistemas de IA los que se empleen para la generación o manipulación de *deepfakes*. Sin embargo, podría darse el caso de modelos de uso general de IA que produzcan *deepfakes*, directamente. Este sería probablemente un caso de modelos de IA de uso general que presentan riesgos sistémicos, en cuyo caso la competencia para actuar sería de la Comisión Europea. Pero como decimos, el reglamento no plantea el supuesto de hecho en el que la *deepfake* se genere directamente por un modelo de IA.

d) el incumplimiento del art. 50».

Para interpretar en toda su amplitud el art. 79.6, nos tenemos que remitir, en primer lugar, al párrafo 5 del art. 79, que establece que la autoridad de vigilancia del mercado⁴⁹ «adoptará todas las medidas provisionales adecuadas para prohibir o restringir la comercialización del sistema de IA en su mercado nacional o su puesta en servicio, [o] para retirar el producto o el sistema de IA [...] o recuperarlo». Dicha autoridad notificará estas medidas sin demora indebida a la Comisión y a los demás Estados Miembros. Esta es la notificación a la que se refiere el art. 79.6. Al mismo tiempo, este párrafo 5 remite al párrafo 2 del art. 79, que otorga facultades a la autoridad de vigilancia del mercado para que evalúe los riesgos que un sistema de IA produce y para que, si como resultado de tal evaluación constata que el sistema de IA no cumple los requisitos y obligaciones establecidos en el reglamento, exija «al operador pertinente» o bien que adopte medidas correctoras oportunas, o bien que lo retire o que lo recupere en un plazo que no podrá ser mayor de 15 días, salvo que la legislación específica de la UE de armonización de productos establezca otra cosa. Finalmente, el art. 79.8 establece que si en el plazo de tres meses desde la recepción de la notificación antes mencionada, ninguna autoridad de vigilancia del mercado de los Estados miembros, ni la propia Comisión Europea, presentan objeción alguna en relación con las medidas provisionales adoptadas por una autoridad de vigilancia del mercado, dichas medidas se entenderán «justificadas», lo que debemos de entender por «consolidadas». Ello se entiende, por supuesto, sin perjuicio de los derechos de recurso que la legislación nacional, y en su caso, comunitaria, establezcan en favor del operador al que fueron dirigidas las medidas correctoras antes señaladas.

Por tanto, y en resumen: es evidente que el art. 79 va dirigido a los sistemas de IA, y no específicamente a los contenidos producidos o generados por determinados sistemas de IA. Esto es un efecto, y un defecto, del abordaje del reglamento basado en los riesgos potenciales y actuales de los sistemas de IA (y no en los contenidos). Es cierto que el art. 79.6 habla del «incumplimiento del art. 50», pero en ningún caso se establecen medidas correctoras específicas dirigidas a contenidos. Es decir, el sistema de IA puede establecer, por ejemplo, una funcionalidad para que el usuario del mismo señale que un determinado contenido es una *deepfake*; pero, ¿qué ocurre si ese usuario no emplea esa aplicación para etiquetar el contenido como *deepfake*, a pesar de que el sistema de IA sí que permite dicha posibilidad? Desde la perspectiva del reglamento, el sistema de IA cumplirá con los requisitos y obligaciones establecidos por el art. 50 (y el resto del reglamento), pero el usuario no estará cumpliendo con lo establecido

⁴⁹ Definida por el reglamento de IA, en su art. 3.26, como «la autoridad nacional que lleva a cabo las actividades y adopta las medidas previstas en el reglamento (UE) 2019/1020», relativo a la vigilancia del mercado y la conformidad de los productos y por el que se modifican la Directiva 2004/42/CE y los Reglamentos (CE) 765/2008 y (UE) 305/2011.

en el art. 50.4. Para estos casos, el reglamento no establece un sistema claro de corrección que pueda aplicarse al usuario del sistema de IA.

¿Se pueden imponer, empero, sanciones y multas administrativas a aquellos operadores que generen *deepfakes* y las introduzcan en las redes? La respuesta que da el reglamento tampoco es completamente clara a este respecto. En efecto, el art. 99 del mismo establece, en su apartado 4 g), que el incumplimiento por parte de proveedores y responsables del despliegue de las obligaciones establecidas en el art. 50 podrán ser sancionables con multas de hasta 15 millones de euros o, si el infractor es una empresa, hasta el 3% de su volumen de negocios mundial total correspondiente al ejercicio financiero anterior (si esta cuantía fuese superior). Por tanto, si bien es cierto que el art. 99 habla del responsable del despliegue, el art. 3.4) define a este sujeto como aquel que usa el sistema de IA «salvo cuando su uso se enmarque en una actividad personal de carácter no profesional». En consecuencia, todos aquellos usuarios de sistemas de IA que puedan generar *deepfakes* cuando lo hagan de manera personal y no profesional no entran dentro del supuesto de hecho que prevé el art. 99.4º. El radio de eficacia jurídica de dicho artículo se limita a los responsables del despliegue de sistemas de IA, tal y como los define el reglamento. Es criticable esta restricción que hace el reglamento, porque en muchos casos, podrán ser precisamente este tipo de usuarios los que introduzcan *deepfakes* en las redes usando sistemas de IA al efecto. Para el resto de los usuarios de sistemas de IA que generen *deepfakes* (es decir, aquellos que lo hagan de una manera profesional) las multas establecidas en el reglamento en el art. 99.4º sí que serán de eventual aplicación.

V. CONCLUSIONES Y CONSIDERACIONES «DE LEGE FERENDA»

Como hemos visto en este trabajo, la regulación de las *deepfakes* en el ámbito comunitario, a través del reglamento de IA, es claramente insuficiente. En primer lugar, el abordaje adoptado por el reglamento, en virtud del cual se regulan riesgos, y no contenidos directamente, genera claras disonancias regulatorias en el ámbito específico de las *deepfakes*. En este sentido, no se ve claramente cómo se podría conseguir que un usuario que no ha etiquetado una *deepfake* como tal lo acabe haciendo, sobre todo si este usuario no es un usuario profesional. El reglamento está pensando en los sistemas de IA (y en su caso, en los modelos de uso general de IA) y no tanto en los productos o resultados de dichos sistemas o modelos. Por lo tanto, es perfectamente posible que dichos sistemas o modelos cumplan perfectamente con los requisitos y obligaciones establecidos por el reglamento, a pesar de lo cual puedan emplearse «for bad», para hacer el mal, por ejemplo, para generar *deepfakes* que atenten contra el sistema democrático. Las soluciones ante este problema que ofrece el reglamento no son por lo tanto claras, tanto en la fase de la prevención, como en la de la corrección, y también en la de

la sanción, puesto que las sanciones parecen pensadas para usuarios profesionales, no para usuarios no profesionales. La definición restrictiva que hace el reglamento de los responsables del despliegue de sistemas de IA es otro problema en este sentido, como hemos visto más arriba. Además, el reglamento no regula de manera adecuada las *deepfakes* de texto, puesto que establece una categoría adicional pero distinta a las *deepfakes* de audio, video e imagen, que no tiene, bajo mi punto de vista, sentido. Parece que el reglamento obliga a etiquetar las manipulaciones de texto cuando se usan para el «bien» (con excepciones), pero se olvida de las manipulaciones de texto cuando se usan para «el mal», cuando no se originan en contenidos sintéticos, lo que es claramente disruptivo desde un punto de vista jurídico. Incluso en el terreno de la generación o manipulación de texto para hacer el bien se plantean problemas, puesto que las obligaciones de información se relajan si el texto generado o manipulado es supervisado por un humano o es fruto de una labor editorial, lo que sobre todo en relación con el primer caso será difícil de probar; además, son excepciones que carecen de sentido, puesto que no vemos de qué manera podría mermar el disfrute de textos generados para hacer el bien la divulgación, a través de un etiquetado, de que las mismas se han realizado de manera artificial. En definitiva: a pesar de que el reglamento es un primer paso en la regulación de las *deepfakes*, es evidente que el mismo no crea las herramientas adecuadas para luchar contra un fenómeno que puede ser empleado para alterar de manera profunda el normal funcionamiento de nuestras democracias.

A la vista del análisis anterior, parece oportuno concluir este trabajo realizando las siguientes consideraciones *de lege ferenda*. Como vemos, y con carácter general, la regulación de las *deepfakes* para proteger la democracia es muy deficiente en el actual sistema establecido por el reglamento de IA. Se proponen, por tanto, las siguientes mejoras:

- 1) En primer lugar, no podemos olvidar que nuestro objeto a proteger son aquellos usuarios de redes sociales o de internet que creen de manera honesta en un contenido que finalmente es una ultrasuplantación, una *deepfake*. Directamente conectado con el anterior, existe otro bien jurídico a proteger, que es la democracia, y más concretamente, la confianza en los sistemas democráticos. Las *deepfakes* erosionan nuestra confianza en los sistemas democráticos, corrompiendo la esfera y el debate públicos, que son consustanciales a la democracia. Aquellas personas que creen honestamente en los contenidos que figuran por internet y por las redes sociales, merecen una protección especial por parte del ordenamiento jurídico.
- 2) Como señalábamos en la primera parte de este trabajo, esas personas, objeto de protección, pueden recibir y a su vez diseminar *deepfakes* porque realmente crean en ellas o porque busquen la confirmación de sus sesgos políticos. Tanto unas como otras merecen protección, pero, como veremos más adelante, la situación del intérprete y aplicador del derecho, en un caso y en otro, no será la misma. Esto tiene que ver con el análisis o estándar contra-

factual que examinaremos más adelante. En cualquier caso, ambos grupos pueden beneficiarse por igual de medidas que amplíen la alfabetización en materia de sistemas y modelos de IA. En este sentido, los Estados deberían promover programas educativos que tengan como finalidad la alfabetización en materia de IA de la población. El reglamento de IA hace alusión a la cuestión de la alfabetización en materia de IA en varias ocasiones (véase concretamente el art. 4 del mismo), pero no establece medidas concretas al respecto. Es urgente que los Estados y la Unión Europea las aborden de manera inmediata

- 3) Es importante que en la definición de *deepfakes* se incluya el texto. Ya hemos visto que la diferenciación que establece el reglamento de IA en materia de ultrasuplantaciones y manipulaciones de texto, en el art. 50.4 del mismo, carece completamente de sentido. Las *deepfakes* pueden tener contenidos textuales y, por lo tanto, se impone una regulación unívoca al respecto.
- 4) La figura sobre la que debería pivotar la responsabilidad en materia de *deepfakes* es el usuario de sistemas de IA. En este sentido, la figura de la persona que «despliega» el sistema de IA es confusa en el reglamento, porque parece que está haciendo mención a una actividad que va más allá que el mero uso de un sistema o modelo de IA para generar *deepfakes*. Además, el hecho de que esa persona esté empleando el sistema o modelo de IA de manera profesional o no profesional es irrelevante desde la perspectiva del bien jurídico que queremos proteger, que es la democracia. Ambos (usuarios profesionales y no profesionales) deberían quedar sujetos a las obligaciones y los límites jurídicos que figuran a continuación.
- 5) La generación de *deepfakes* que tenga un impacto directo en la democracia, en la salud de nuestro debate público, y en la esfera pública, deberían quedar no solamente sometidas a obligaciones de etiquetado, sino directamente prohibidas por nuestros sistemas jurídicos. Se debería establecer un sistema de sanciones y multas, en su caso de tipo penal, para aquellos usuarios de sistemas y modelos de IA que las introduzcan y las difunden, a sabiendas, en las redes sociales.
- 6) En los demás casos (es decir, *deepfakes* que puedan tener un impacto indirecto en la democracia, y, en general, todas las *deepfakes*) deberían quedar sometidas a claras obligaciones de etiquetado. Se propone que se siga a este respecto el modelo californiano⁵⁰. Por tanto, las personas que introduzcan

⁵⁰ En efecto, California ha regulado las *deepfakes* empleadas en período electoral. Concretamente, la Ley AB-2655 exige que las plataformas de gran tamaño como Facebook o X (Twitter) establezcan mecanismos para denunciar *deepfakes* electorales y que esas plataformas eliminen o al menos etiqueten dichos *deepfakes*. La regulación californiana es tan precisa que llega incluso a establecer lo que la etiqueta debe indicar. Por su parte, la Ley AB-2839 regula la acción o acciones de los usuarios que publican en la red o republican (retuitean, por ejemplo) *deepfakes* que pudieran inducir a error a los votantes, imputándoles responsabilidad civil por diseminar falsa información. Es

- deepfakes* en las redes sociales y en internet deberían incluir una etiqueta que dijera: «Este contenido es una ultrasuplantación». Se deberían aprobar normas de hard-law en donde se estableciera cómo debe ser dicha etiqueta, donde se debe «pegar», cuál debe ser su tamaño de letra, etc. No es suficiente, en este sentido, con que las redes sociales de gran tamaño establezcan funcionalidades para que los usuarios puedan etiquetar las *deepfakes*: hay que obligar, directamente, a dichos usuarios a que realicen el etiquetado.
- 7) Se debe reconocer que tanto los modelos de IA como los sistemas de IA son capaces de generar *deepfakes*. Por tanto, la regulación sobre la materia debería contemplar ambas posibilidades, por mucho que la regla habitual sea que la generación de *deepfakes* provenga de sistemas de IA y no de modelos de IA. En este sentido, la futura regulación de las *deepfakes* deberían atender al contenido, y no a la capacidad o no de sistemas y modelos de IA de generar determinados riesgos. Existen sistemas y modelos de IA, como, por ejemplo, ChatGPT, que no están específicamente destinados a generar *deepfakes*, y que, sin embargo, como hemos visto en este artículo, son capaces de hacerlo. Insisto: el acento regulatorio debe ponerse en los contenidos y no en los riesgos, al menos en materia de *deepfakes*.
 - 8) Adicionalmente, se deben establecer sistemas específicos de corrección de aquellas *deepfakes* que no están etiquetadas correctamente, y, en su caso, de retirada de las mismas. Se debería atribuir competencia para ambas acciones a las Administraciones competentes en materia de IA tanto en el nivel comunitario como en el nivel nacional.
 - 9) Además, se deberían desarrollar sistemas específicos de responsabilidad civil y administrativa como consecuencia del daño producido por la generación y difusión de *deepfakes*.
 - 10) Con respecto a los tribunales de justicia, se propone un análisis contrafactual para examinar los asuntos que tengan que decidir en materia de *deepfakes*⁵¹. La pregunta contrafactual que debe plantearse el juez es la de cómo

decir, la persona afectada por la *deepfake* está facultada para ejercitar acciones indemnizatorias contra el usuario que publicó o republicó la *deepfake*. Estas obligaciones de no hacer que impone la ley quedan relajadas en el caso en el que se etiquetara la *deepfake* como tal, pero, de nuevo, la ley en cuestión señala cómo, exactamente, se tiene que etiquetar la misma. Finalmente, la Ley AB-2355 establece que la publicidad electoral que se realice empleando *deepfakes* lo indique de manera clara, de tal manera que se asegure la transparencia en las campañas políticas. De nuevo, la ley californiana es precisa y exhaustiva en lo que se refiere a cómo se tiene que divulgar, exactamente, que la publicidad electoral ha sido realizada por IA y constituye una *deepfake*. Véase con carácter general, Ch. Howell y L. Allen (2024), «Decoding California's Recent Flurry of AI Laws». October 2024) *Journal of Intellectual Property Law & Practice*, 15, 1, págs. 24-31. <https://is.gd/BNu0jU>.

⁵¹ Véase con carácter general mi trabajo: A. Estella (2016), «Rationality and Counterfactual Legal Analysis. *Italian Law Journal*, 2, 1, págs. 105-129.

quedaría la democracia, el debate público, y la esfera pública, en ausencia de una determinada *deepfake*. Si esa *deepfake*, si ese contenido, no se hubiera introducido en el sistema, ¿estaría la democracia peor, mejor o igual? En el caso en el que la respuesta fuera la primera, el tribunal debería tomar la decisión de prohibir la *deepfake* y extraer todas las consecuencias derivadas de esa decisión, sobre todo en materia de responsabilidad civil, administrativa y penal. En este sentido, pesará en la elaboración del contrafactual el hecho de que el usuario (por ejemplo, que diseminó una *deepfake*) haya buscado la confirmación de un sesgo político o que haya simplemente creído en la *deepfake* falsa. Evidentemente, en el primer caso, el contrafactual (que no hubiera habido *deepfake*) adoptará una forma más similar al caso de origen (que haya habido *deepfake*); mientras que, en el segundo caso, la distancia entre el contrafactual y la situación fáctica será mayor. Esta distancia (como decimos, mayor en el segundo caso, menor en el primero) será tenida en cuenta por el juez a los efectos de la valoración de las consecuencias jurídicas que deben conectarse a la inserción y diseminación de la *deepfake* en internet o en redes sociales. Es decir: frente al parámetro subjetivo de la intención del usuario del sistema o modelo de IA (generar o no una disrupción en la democracia) o el estándar objetivo del daño que actual o potencialmente se hubiera podido crear, este trabajo propone un tercer estándar que es el del análisis de la situación contrafactual, en la que el juez debe comparar la situación fáctica que realmente se ha producido con la que se habría podido producir en ausencia de la *deepfake*.

