Mariona Cardona Valles
Patricia Hernández Hidalgo
Miquel Peguera Poch
Anna María Ruiz Martín
Editores

Desafíos actuales de la Inteligencia Artificial

Actas del XIX Congreso IDP 2024 (Internet, Derecho y Política)

Universitat Oberta de Catalunya

Edicions MIC

ISBN:
Impreso en España

PRESENTACIÓN

La inteligencia artificial (IA) ha emergido como elemento disruptivo de primera magnitud. Es un motor fundamental de crecimiento en todos los sectores económicos y está transformando la forma en que vivimos, nos comunicamos, trabajamos y consumimos. Su capacidad para procesar grandes cantidades de datos ha impulsado avances significativos en la medicina, la industria, la educación y otros muchos campos. Ha fomentado la automatización, ahorrando tareas repetitivas y permitiendo un enfoque en actividades más creativas. Se esperan de ella resultados relevantes para abordar desafíos globales como el cambio climático, la lucha contra enfermedades raras o la obtención de nuevas fuentes de energía.

A la vez, la IA plantea riesgos innegables. La opacidad de los algoritmos, los sesgos incorporados en los datos de entrenamiento, y las decisiones automatizadas basadas en estos sistemas pueden ocasionar resultados discriminatorios en múltiples esferas, como el acceso a servicios esenciales, la obtención de crédito, la contratación de seguros de vida y salud, el disfrute de prestaciones públicas, o la inserción y promoción laboral, entre otros muchos campos. La capacidad intrusiva de los sistemas de reconocimiento facial, o de sistemas de IA de inferencia de emociones y características sensibles resulta difícilmente compaginable con los derechos a la intimidad y a la protección de datos. La capacidad de generar contenidos manipulados y expandir la desinformación plantea desafíos al funcionamiento democrático y al derecho a la información. Estos y otros retos desafían a la ciudadanía, a la industria, y también a los Estados, que deben decidir si regulan la IA, y de qué modo hacerlo.

La Unión Europea es ya pionera en la regulación de la Inteligencia Artificial gracias al Reglamento de IA publicado el 12 de julio de 2024, que da un gran paso en la creación de un marco de seguridad jurídica para la adopción de una IA fiable y centrada en el ser humano y para mitigar los riesgos graves contra la salud, la seguridad y los derechos fundamentales. En otras jurisdicciones están avanzando también trabajos legislativos relevantes. Por otra parte, la comunidad académica está siguiendo de cerca estas iniciativas reguladoras así como el desarrollo vertiginoso de la IA desde ángulos muy diversos.

El XIX Congreso IDP (Internet, Derecho y Política), con el título de Desafíos actuales de la Inteligencia Artificial, impulsado por los Estudios de Derecho y Ciencia Política de la Universitat Oberta de Catalunya y celebrado en la misma universidad el 17 de octubre de 2024, ha tenido por objeto abordar aspectos clave de la IA desde las perspectivas jurídica, politológica, criminológica y de las relaciones internacionales. El presente libro recoge las comunicaciones enviadas y aceptadas, la mayoría de las cuales fueron objeto de presentación oral y discusión en las sesiones del congreso. Con esta publicación, disponible en acceso abierto, esperamos contribuir al necesario debate en torno a los desafíos que la IA sigue planteando.

Barcelona, octubre de 2024.

Capítulo 1

REGLAMENTO DE INTELIGENCIA ARTIFICIAL, DIRECTIVA DE PLATAFORMAS Y LEYES NACIONALES SOBRE GESTIÓN ALGORÍTMICA: SUPERPOSICIÓN Y VACÍOS DE PROTECCIÓN[1]

Maria del Mar Crespí Ferriol
Profesora Permanente Laboral de Derecho del Trabajo
Universidad de las Islas Baleares

RESUMEN: En este trabajo se analizan los distintos ámbitos de aplicación y niveles de protección que establecen las normas europeas que afectan a la gestión algorítmica en el ámbito laboral. A partir de su comparación, se concluye que ninguna norma ofrece una garantía satisfactoria de los derechos de los trabajadores, si se entiende como tal el establecimiento de (i) derechos de transparencia individual, (ii) derechos colectivos de información y consulta de los representantes de los trabajadores y (iii) limitaciones específicas del uso de los sistemas informáticos. El Reglamento de Inteligencia Artificial solo se aplica a tecnologías muy avanzadas, la Directiva de Plataformas es una norma de carácter sectorial y otras normas como el Reglamento de protección de datos solo ofrecen una protección parcial. Por su parte, las normativas nacionales que han adoptado hasta ahora los estados miembros presentan problemas similares. Las leyes nacionales que contienen una regulación exhaustiva la restringen a las plataformas digitales y las que son de aplicación general a todos los trabajadores tienen un contenido limitado. Por tanto, en ambos niveles normativos se genera un vacío de protección para los trabajadores que no prestan servicios para plataformas digitales y que pueden estar afectados por sistemas de gestión algorítmica que no merezcan la calificación de sistemas de inteligencia artificial. Los legisladores nacionales tendrán que decidir si cubrir dicho vacío a la hora de transponer la Directiva de Plataformas extendiendo su ámbito de aplicación personal a todos los trabajadores.

PALABRAS CLAVE: Gestión Algorítmica, Inteligencia Artificial, Plataformas Digitales de Trabajo, Principio de Transparencia, Protección de datos

1 El presente trabajo forma parte del Proyecto de Investigación INTELIGENCIA ARTIFICIAL Y DERECHO: ANALISIS DE LA RESPONSABILIDAD DE LOS DAÑOS DERIVADOS DEL USO DE SISTEMAS DE INTELIGENCIA ARTIFICIAL, PID 2022-140944OA-100 financiado por MCIN/AEI/10.13039/501100011033/FEDER, UE. Su elaboración es fruto de la estancia de investigación realizada en la Fundación Europea para la Mejora de las Condiciones de Vida y de Trabajo (Eurofound), en Dublín (Irlanda). Por ello, deseo mostrar mi agradecimiento a Ricardo Rodríguez por su valiosa ayuda y apoyo en el desarrollo en esta idea y al resto de los investigadores de las Unidades A y B por inspirar la investigación en esta materia.

1. INTRODUCCIÓN

En los últimos años, las instituciones europeas han venido realizando un gran esfuerzo legislativo dirigido a afrontar los retos que la transformación digital comporta para la sociedad. Más concretamente, en el ámbito laboral, se han abordado los retos que las tecnologías basadas en el tratamiento informático avanzado de datos personales suponen para la garantía y la efectividad de los derechos de los trabajadores. Con tal fin, la Unión Europea ha adoptado regulaciones tecnológicas transversales que incorporan ciertas disposiciones con algún contenido laboral, como el Reglamento General de Protección de Datos[2] (RGPD) o el Reglamento de Inteligencia Artificial[3] (RIA). Y, a la vez, ha adoptado normas de fundamentación laboral, que tienen un impacto directo en la aplicación de tecnologías de gestión algorítmica de los recursos humanos. La más reciente y relevante es la Directiva sobre la mejora de las condiciones laborales de los trabajadores de plataformas (DTP), pero también tienen relevancia en esta materia la Directiva de Condiciones Transparentes[4] o la más antigua Directiva de Información y Consulta de los Representantes de los Trabajadores[5]. Como consecuencia de todo ello, en el ámbito de la gestión algorítmica se produce una acumulación de fuentes europeas que se han ido superponiendo unas a otras dando resultado a una elevada complejidad normativa[6].

Pues bien, a esta lista no exhaustiva de normas de origen europeo, se suman las iniciativas legales que han ido adoptando los distintos estados miembros, con perspectivas y orientaciones muy distintas. Algunas de ellas son resultado de la transposición o ampliación de las normas europeas ya mencionadas, como la Directiva de condiciones transparentes o el RGPD y otras las concibieron los estados por iniciativa propia[7], ya fuera con un propósito explícitamente laboral o enmarcadas dentro de un programa normativo más amplio. Esto implica que la multiplicidad de obligaciones empresariales que impone la normativa europea se solapa también, en cierto modo, con las obligaciones que ya han venido fijando por su cuenta los estados miembros.

2 Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE.

3 Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) nº 300/2008, (UE) nº 167/2013, (UE) nº 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.

4 Directiva (UE) 2019/1152 del Parlamento Europeo y del Consejo, de 20 de junio de 2019, relativa a unas condiciones laborales transparentes y previsibles en la Unión Europea.

5 Directiva 2002/14/CE del Parlamento Europeo y del Consejo, de 11 de marzo de 2002, por la que se establece un marco general relativo a la información y a la consulta de los trabajadores en la Comunidad Europea.

6 CODAGNONE, Cristiano; WEIGL, Linda: "Leading the Charge on Digital Regulation: The More the Better, or Policy Bubble?", *Digital Society*, (2023), 2(4).

7 EUROFOUND, Regulatory responses to algorithmic management in the EU, 2024, https://www.eurofound.europa.eu/en/resources/article/2024/regulatory-responses-algorithmic-management-eu

Toda esta producción normativa deriva en algunos efectos jurídicos paradójicos. En primer lugar, las normas mencionadas convergen y divergen en distintos aspectos cuando delimitan sus ámbitos de aplicación material y personal. Ello significa que la gestión algorítmica se regirá por parámetros jurídicos distintos dependiendo de la tipología de tecnología que se utilice en cada caso, así como de las empresas que la implanten. En segundo lugar, las normas europeas y nacionales fijan niveles muy distintos de protección, aunque en el fondo circundan, en todo o en parte, en torno a unas mismas ideas generales: (i) la fijación de obligaciones que concretan el principio de transparencia individual, (ii) el establecimiento de deberes de información y consulta de los representantes de los trabajadores y (iii) la prohibición de usos específicos de los medios tecnológicos con el fin de proteger derechos fundamentales o laborales particulares. Todo ello implica que, por un lado, en los ámbitos en los que coincidan distintas normas habrá empresas a las que se apliquen múltiples obligaciones similares, pero no necesariamente coincidentes, pudiéndose generar duplicidades o confusiones innecesarias. Y, por otro lado, implica que en los ámbitos en los que no coincidan dichos planos normativos solo se apliquen regulaciones con una relación tangencial con el hecho tecnológico, por lo que habrá trabajadores que puedan quedar total o relativamente desprotegidos.

Este trabajo se propone realizar una aproximación general a dicho problema en un momento clave de la transformación digital y jurídica relacionada con la gestión algorítmica. Así, el RIA prevé una aplicabilidad progresiva a lo largo de los próximos tres años, según sus distintas materias (art. 113). E, igualmente, es inminente que empiece a contar el plazo de dos años que la DTP otorga a los estados miembros, desde el momento de su publicación, para adoptar las leyes, reglamentos y disposiciones administrativas necesarias para cumplirla (art. 29.1). Con dicho propósito se realiza una revisión general de la citada legislación europea, incorporando también el análisis de las leyes nacionales de una selección de nueve estados miembros. Para la dimensión nacional del estudio, se utilizan la Platform Economy Database[8] y la base de datos legal sobre gestión algorítmica del European Research Monitor[9] de Eurofound. Ambos recursos contienen información actualizada sobre las iniciativas legales o de políticas públicas que los estados han planteado en relación con la transformación digital del mundo laboral. A partir de ahí, se analiza el ámbito de aplicación de las normativas relacionadas con la protección de los trabajadores frente a la gestión algorítmica, distinguiendo su aspecto material y personal. A continuación, se lleva a cabo una evaluación general del contenido y el nivel de protección que ofrece cada norma a partir de su atención a nueve ítems jurídicos concretos. Por último, se realizan algunas propuestas para orientar la futura transformación de los ordenamientos jurídicos nacionales.

8 EUROFOUND, Platform Economy Database, (2023), Dublin, https://eurofound.link/platformeconomydb

9 EUROFOUND: Algorithmic management, Restructuring legislation database, (2024), Dublin, https://apps.eurofound.europa.eu/legislationdb/algorithmic-management

2. LOS ÁMBITOS DE APLICACIÓN DE LAS NORMATIVAS EUROPEAS Y NACIONALES SOBRE GESTIÓN ALGORÍTMICA

Cuando se afirma que en el ámbito de la gestión algorítmica se está produciendo una superposición o reiteración normativa ello significa que existe una pluralidad de normas que fijan iguales o similares obligaciones a los mismos sujetos que se encuentran en las mismas circunstancias. Por ello, antes de entrar a analizar en concreto las obligaciones que imponen las normas europeas y nacionales, es necesario detenerse en aclarar cuando son de aplicación.

2.1. El ámbito de aplicación material

El ámbito de aplicación material de las normas que tratan sobre la gestión algorítmica se define principalmente por el tipo de tecnología sobre las que se proyectan.

En primer lugar, cualquier clase de gestión digital de datos personales relativos a los trabajadores debe sujetarse a los parámetros del RGPD, que son aplicables a cualquier tecnología en un sentido genérico. Si bien, el propio RGPD contempla algunas reglas más estrictas para determinados sistemas más sofisticados de tratamiento de datos personales, como el perfilado o los sistemas de decisión individual automatizada (SDIA). Estos últimos se definen en el artículo 22 RGPD como sistemas que toman decisiones basadas únicamente en el tratamiento automatizado de datos, incluida la elaboración de perfiles, que producen efectos jurídicos en una persona o le afectan significativamente de modo similar.

En segundo lugar, se encontrarían los sistemas de gestión algorítmica. En general, se habla de gestión algorítmica para referirse a prácticas en las que se usan datos personales de los trabajadores o sobre el proceso de trabajo para alimentar algoritmos, los algoritmos procesan dichos datos y los analizan, y estos dos elementos apoyan la coordinación y control de los trabajadores que ejerce la dirección empresarial[10]. La DTP, a su vez, distingue dos tipologías de gestión algorítmica: los sistemas automatizados de control y los sistemas automatizados de toma de decisiones. La primera se refiere a sistemas que se utilizan para o que apoyan el seguimiento, la supervisión o la evaluación del rendimiento laboral o de las actividades realizadas en el entorno de trabajo (art. 2.8). La segunda, a sistemas que se utilizan para tomar o apoyar, mediante medios electrónicos, decisiones que afectan significativamente a los trabajadores, incluidas sus condiciones de trabajo (art. 2.9).

Como se ve, tanto el RGPD como la DTP solo aplican sus requerimientos específicos a sistemas informáticos cuyas decisiones afecten a los trabajadores de manera 'singificativa', siendo este un término que puede dar lugar a dificultades interpretativas. No obstante, la DTP adopta una perspectiva más amplia al abarcar tanto los sistemas usados para tomar

10 BAIOCCO, Sara; FERNÁNDEZ-MACÍAS, Enrique, RANI, Uma; PESOLE, Annarosa: *The Algorithmic Management of work and its implications in different contexts*, European Commission, JRC Working Papers Series on Labour, Education and Technology (2022).

directamente dichas decisiones, como los usados para dar soporte a decisiones tomadas por humanos. De tal modo, se consigue ampliar la protección legal a aquellos supuestos en los que la tecnología se utiliza de modo auxiliar y, además, se impide que esta pueda evadirse con una intervención humana formal o meramente testimonial[11].

La tipología más específica es aquella que se refiere a los sistemas de inteligencia artificial (IA), tal y como aparecen definidos en el artículo 3.1 RIA. Son sistemas basados en máquinas diseñados para operar con distintos niveles de autonomía, que pueden exhibir adaptabilidad después de su implementación y que, con fines explícitos o implícitos, infieren, a partir de los inputs que reciben, cómo generar resultados como predicciones, contenido, recomendaciones o decisiones que puedan influir en entornos físicos o virtuales. Este concepto va más allá de lo que se había previsto en el Proyecto inicial de 2021 del RIA[12], que se había criticado por excesivamente amplio[13]. Ahora, la IA debe distinguirse de sistemas informáticos tradicionales más simples u otros enfoques de programación a través de una serie de características (autonomía, inferencia, adaptabilidad) que aparecen definidas en el Considerando 12[14] y son las que, a su vez, permiten diferenciarla de otras formas de gestión algorítmica. Es decir, el uso de sistemas de IA para fines laborales puede considerarse un subtipo de gestión algorítmica, pero no todos los programas informáticos que usen algoritmos son lo suficientemente avanzados como para considerarse sistemas de IA.

2.2. El ámbito de aplicación personal

La segunda variable que determina la relevancia de las normas que se analizan es el ámbito de aplicación personal. En primer lugar, las normas de alcance más amplio son las que tienen una vocación transversal y, por lo tanto, se proyectan más allá del ámbito laboral, aun cuanto contemplan en cierto modo su particularidad. Este es el caso del RGPD (art. 88) y del RIA (art. 26.7). En segundo lugar, se encuentran las normas de ámbito estrictamente

11 ALOISI, Antonio; POTOCKA-SIONEK, Nastazja: *De-gigging the labor market? An analysis of the 'algorithmic management' provisions in the proposed Platform Work Directive*, Italian Labour Law e-Journal, (2022), 15(1), pp. 30-50.

12 Un sistema de IA se definía como "software que se desarrolla con una o más de las técnicas y enfoques enumerados en el Anexo I y que puede, para un conjunto determinado de objetivos definidos por humanos, generar resultados como contenido, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa".

13 RUSCHEMEIER, Hannah: "AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal", *ERA Forum* (2023), 23, pp. 361-376.

14 La primera de estas características es que los sistemas de IA deben operar con autonomía, "lo que significa que tienen un cierto grado de independencia de las acciones del ser humano y capacidades para operar sin intervención humana". La segunda es que deben ser capaces de inferir o de obtener resultados, así como de configurar algoritmos a partir de los propios datos que se incorporan al sistema. Es decir, no se limitan a aplicar reglas predefinidas de forma automatizada, sino que son capaces de generar sus propias pautas a partir de técnicas de aprendizaje automático con capacidad de razonar. La tercera característica distintiva es que son sistemas adaptables, que evolucionan por sí mismos después de su instalación y durante su uso continuado.

laboral, concebidas para ser aplicadas a la relación de trabajo subordinado. Este es el caso de la mayoría de las directivas europeas en materia de política social, como la Directiva de condiciones transparentes o la Directiva de información y consulta. En tercer lugar, con carácter más restringido, pueden distinguirse las normas laborales de aplicación sectorial, diseñadas para adaptarse a las particularidades de las empresas que presentan un modelo económico similar, como ocurre con la DTP. No obstante, uno de los aspectos más novedosos de esta Directiva es que extiende parte de sus efectos fuera del trabajo subordinado. Así, parte de sus disposiciones se aplican también a lo que denomina personas que prestan servicios para las plataformas (art. 1.2 DTP), como categoría que incluye no solo trabajadores con un contrato laboral, sino también trabajadores autónomos.

Como se ve, la legislación de la Unión Europea conectada con la gestión algorítmica (i) incluye previsiones de orientación laboral en normativas de ámbito transversal y (ii) fija una legislación específicamente laboral de la gestión algorítmica, que, sin embargo, se limita al sector de las plataformas de trabajo. Ello implica que la gestión algorítmica carece de una legislación laboral que pueda aplicarse a otros ámbitos de la actividad económica. Dicho vacío de protección debe cubrirse, en la medida de lo posible, con las reglas de orientación laboral que contienen las normas transversales, que solo ofrecen una protección parcial y fragmentaria, y con la aplicación Directivas laborales, que no contemplan específicamente el factor tecnológico.

2.3. El ámbito de aplicación de las normas nacionales

En el siguiente cuadro-resumen se refleja el ámbito de aplicación de algunas normativas nacionales[15] relacionadas con la gestión algorítmica, comparándolas con las legislaciones europeas que ya se han descrito. Como queda reflejado, las legislaciones estatales son muy diversas en sus ámbitos de aplicación. Si bien, en cuanto al ámbito de aplicación material, puede afirmarse que la tendencia mayoritaria consiste en establecer regulaciones que abarcan las tecnologías de gestión algorítmica en un sentido amplio. El hecho de que, por ejemplo, la legislación italiana se aplique únicamente a las decisiones automatizadas sin intervención humana ha sido objeto de crítica por parte de los sindicatos, que entienden que ofrece una protección insuficiente a los trabajadores[16]. La legislación búlgara es la más específica porque se aplica únicamente a la utilización de algoritmos para el control del tiempo de trabajo de teletrabajadores. Por tanto, no es solo que su ámbito de aplicación personal sea muy restringi-

15 De entre los 12 países que se contemplan en la base de datos del European Restructuring Monitor sobre gestión algorítmica, Noruega y Finlandia se dejan fuera del presente estudio, porque no tienen una regulación tecnológica especifica. Dicha base de datos se refiere a sus legislaciones laborales generales que podrían aplicarse a la introducción de la gestión algorítmica, como, por ejemplo, la obligación de informar a los trabajadores antes de introducir nuevos medios tecnológicos de control, pero no tratan la gestión algorítmica en particular.

16 EUFOFOUND, Italy: Algorithmic management, Restructuring legislation database, (2024) Dublin, https://apps.eurofound.europa.eu/legislationdb/algorithmic-management/italy

do, sino que además se añade un elemento teleológico que lo limita todavía más. Por último, la griega es una legislación transversal, como el RIA, que se aplica con carácter general, pero solo a la subcategoría de tecnologías que merecen la calificación de sistemas de inteligencia artificial.

	Gestión algorítmica	**SDIA**	**SDIA para fines particulares**	**Sistemas de IA**
Aplicación general	*RGPD*			*RIA* Grecia
Trabajadores en general	Alemania Portugal España	*RGPD (art. 22)* Italia		
Trabajadores de plataformas	*DTP* Croacia Francia Malta			
Trabajadores a distancia			Bulgaria	

En cuanto al ámbito de aplicación personal, con las excepciones de Grecia y Bulgaria, los estados miembros se dividen en dos grandes grupos: los que establecen una regulación laboral de carácter general y los que han adoptado leyes destinadas a los trabajadores de plataformas. En particular, el ámbito de aplicación de las legislaciones nacionales de España (Ley Rider) y Portugal (Ley de Trabajo Decente) puede llevar a confusión, porque son legislaciones que se anunciaron con la vocación de dar una respuesta adecuada a los problemas a los que se enfrentaban los trabajadores de plataformas. Sin embargo, se trata de actos legislativos que modificaron en ambos casos el contenido de los respectivos Estatuto de los Trabajadores y Código de Trabalho, por lo que, una vez en vigor, benefician a todos los trabajadores subordinados, con independencia del sector económico al que se dediquen.

Por último, es interesante tener en cuenta que, cuando se hace referencia a las legislaciones abarcan en general a todos los trabajadores, ello incluye no solo a personas que tienen un contrato de trabajo como tal, sino que también contemplan expresamente a las que se encuentran en un proceso de selección. Es el caso, por ejemplo, de Portugal, que coincide en este punto con la DTP (art. 7.2) o el RIA (art. 4.a del Anexo III).

3. DERECHOS O GARANTÍAS JURÍDICAS FRENTE A LA GESTIÓN ALGORÍTMICA

El siguiente cuadro resume los principios y derechos relacionados con la gestión algorítmica que contemplan las distintas legislaciones europeas y nacionales. Se trata de una representación necesariamente simplificada para poder trazar una comparación general que no exceda los límites del propósito de este trabajo. Por lo tanto, se agrupan bajo una misma categoría facultades u obligaciones que pueden estar configuradas de formas muy distintas, con alcances variados en lo que respecta a su presupuesto de hecho, contenido, función o formalidades a los que están sujetos. Por ejemplo, en el caso del principio individual de transparencia, este se concreta, entre otras, en la obligación genérica de informar a los trabajadores afectados de que van a estar expuestos a un sistema de inteligencia artificial (art. 26.7 RIA), en una obligación de contenido mucho más detallado de informar a los trabajadores por escrito sobre todas las decisiones que se apoyan en el uso de sistemas algorítmicos (art. 9 DTP) o en la obligación de informar sobre la lógica del algoritmo implicado en el SDIA, así como las implicaciones y posibles consecuencias de tal procesamiento de datos para el interesado (art. 14.2.g RGPD).

Lo mismo ocurre con las prohibiciones de uso de la tecnología establecidas para proteger derechos específicos de los trabajadores, que se concretan también de formas diversas. Así, por ejemplo, tanto el RIA (art. 5.1.f) como la DTP (art. 7.1.a) prohíben el uso de sistemas de reconocimiento de emociones en el ámbito laboral con el propósito de proteger la intimidad de los trabajadores. Por su parte, tanto la DTP (art. 7.1.c) como la normativa de Malta prohíben el tratamiento de datos personales de los trabajadores que no estén relacionados con su actividad laboral. O, como muestra adicional, en el caso de Francia, se impide que los algoritmos penalicen el ejercicio del derecho de desconexión digital o que se perjudique a los trabajadores que rechacen encargos ofreciéndoles servicios peor remunerados.

A nivel europeo, la norma que regula de forma más completa y exhaustiva la gestión algorítmica es la DTP. Ello se debe a que el modelo de negocio de las plataformas de trabajo se fundamenta precisamente en su base tecnológica y en la gestión algorítmica de las relaciones con sus trabajadores y clientes. Y también, como ya se ha mencionado, a que la DTP es la única norma europea de orientación explícitamente laboral que contempla las particularidades de la gestión algorítmica. Del mismo modo, parece que este patrón se reproduce también a nivel nacional. Así, los estados miembros como Francia, Malta o Croacia, cuyas regulaciones se enfocan en las plataformas digitales de trabajo, presentan una normativa que aborda de forma más exhaustiva y completa el fenómeno de la gestión algorítmica. En cambio, los que presentan regulaciones de alcance general, como Alemania o España, tienen un contenido más restringido, porque se centran eminentemente en el reconocimiento de derechos colectivos de información y consulta de los representantes de los trabajadores.

	DTP	RIA	Art. 22 RGPD	Bulgaria	Croacia	Francia	Alemania	Grecia	Italia	Malta	Portugal	España
Derechos de información individual del trabajador	X	X	X	X	X	X		X	X	X	X	
Obligación empresarial de realizar autoevaluación de riesgos	X	X	X		X							
Obligación empresarial de supervisión humana de SDIA	X		X	X	X							
Derecho colectivo de información de representantes de trabajadores	X	X				X	X		X			X
Derecho sindical de asesoramiento experto	X					X	X					
Información a la autoridad competente	X	X							X	X		
Prohibición de discriminación	X	X	X			X		X			X	
Protección de derechos específicos	X	X			X	X				X		

4. CONCLUSIONES

Las dos normas que contienen una regulación más completa de la gestión algorítmica de los trabajadores son la DTP y el RIA, porque la protección laboral sustantiva que ofrece el RGPD es más débil, aunque tenga un alcance más amplio. Si bien, ni la DTP ni el RIA presenta un tratamiento enteramente satisfactorio de este fenómeno. En el primer caso, porque se trata de una regulación sectorial que presenta un ámbito de aplicación personal restringido a las empresas basadas en plataformas digitales. En el segundo caso porque, pese a ser de cumplimiento obligatorio para todas las empresas, el ámbito de aplicación material del RIA está intencionadamente restringido a sistemas informáticos especialmente sofisticados, que no abarcan todos los programas que pueden utilizarse para la gestión algorítmica.

Pues bien, en el presente trabajo ha podido comprobarse que esta problemática, que ya ha venido siendo señalada por la doctrina a nivel europeo[17], se reproduce también a nivel nacional. Como ha podido verse, las normativas estatales que ya trataban la gestión algorítmica antes de la entrada en vigor del RIA y de la DTP pueden dividirse en dos grupos, a los que se asocian distintos inconvenientes. Las normas nacionales analizadas que presentan ámbitos de aplicación de alcance general a todos los trabajadores y empresas contienen una regulación sustantiva limitada, que normalmente solo contempla deberes parciales de transparencia. En cambio, las normas nacionales que contienen una regulación completa y exhaustiva que contempla derechos individuales y colectivos de los trabajadores, así como garantías para su efectividad, son normas dirigidas exclusivamente a las plataformas digitales de trabajo. Si bien, el uso de algoritmos para gestionar y controlar a los trabajadores ya está extendido en otros sectores económicos, por ejemplo, para la gestión del teletrabajo o para el control de trabajadores a través del internet de las cosas, entre otros usos[18]. Por tanto, de momento las leyes estatales no cubren, sino que más bien reproducen, el vacío de protección que se genera a nivel europeo.

Otro aspecto que pone de manifiesto la comparativa realizada es que, incluso los estados miembros que ya cuentan con leyes dedicadas a las plataformas digitales de trabajo con una regulación relativamente detallada deberán adoptar nuevas reformas para trasponer las disposiciones de la DTP en materia de gestión algorítmica. Es decir, no hay ninguna norma nacional analizada que en estos momentos sea tan avanzada como la DTP. Y, en cierto modo, también deberá concretarse el impacto laboral del RIA, que deja la materialización del deber de transparencia en el ámbito laboral, en parte, a merced de la ley y las prácticas de cada estado miembro (art. 26.7).

Todo lo anterior implica que tal vez la decisión más importante que deberá adoptarse a la hora de trasponer la DTP al ordenamiento interno de los estados miembros será determinar su ámbito de aplicación personal. En España, por ejemplo, habrá que decidir si la nueva

17 ALOISI, Antonio; POTOCKA-SIONEK, Nastazja: De-gigging… cit.

18 DE STEFANO, Valerio; WOUTERS, Mathias: *AI and digital tools in workplace management and evaluation: An assessment of the EU's legal framework*, Brusels, European Parliamentary Research Service, 2022, p. 16.

regulación de la gestión algorítmica se circunscribe a las plataformas de trabajo, como ya ocurría con la presunción de laboralidad introducida por la ley Rider en la Disposición adicional 23 ET, o si se incorpora al texto general del estatuto de los trabajadores, como se hizo con el deber de transparencia algorítmica del art. 64.4.d ET. Esta última opción va más allá de lo que requiere la DTP, pero se cohonestaría bien con sus objetivos y los del RIA, porque la primera solo fija normas de protección mínima de los trabajadores (art. 26) y el segundo contiene lo que se ha denominado principio de no menoscabo de los derechos e intereses de los trabajadores por cuenta ajena[19] (art. 2.11).

5. BIBLIOGRAFÍA

ALOISI, Antonio; POTOCKA-SIONEK, Nastazja: "De-gigging the labor market? An analysis of the 'algorithmic management' provisions in the proposed Platform Work Directive", *Italian Labour Law e-Journal*, (2022), 15(1).

BAIOCCO, Sara; FERNÁNDEZ-MACÍAS, Enrique, RANI, Uma; PESOLE, Annarosa: *The Algorithmic Management of work and its implications in different contexts*, European Commission, JRC Working Papers Series on Labour, Education and Technology (2022).

CODAGNONE, Cristiano; WEIGL, Linda: "Leading the Charge on Digital Regulation: The More the Better, or Policy Bubble?", *Digital Society*, (2023), 2(4).

DE STEFANO, Valerio; WOUTERS, Mathias: *AI and digital tools in workplace management and evaluation: An assessment of the EU's legal framework*, Brusels, European Parliamentary Research Service, 2022.

EUFOFOUND, Italy: Algorithmic management, Restructuring legislation database, (2024) Dublin, https://apps.eurofound.europa.eu/legislationdb/algorithmic-management/italy

EUROFOUND, Regulatory responses to algorithmic management in the EU, 2024, https://www.eurofound.europa.eu/en/resources/article/2024/regulatory-responses-algorithmic-management-eu

EUROFOUND: Algorithmic management, Restructuring legislation database, (2024), Dublin, https://apps.eurofound.europa.eu/legislationdb/algorithmic-management

MERCADER UGUINA, Jesús R.: "El Reglamento de inteligencia artificial entra en la recta final, una primera lectura en clave laboral", *Revista General de Derecho del Trabajo y de la Seguridad Social*, (2024), 67, pp. 327-351.

RUSCHEMEIER, Hannah: "AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal", *ERA Forum* (2023), 23, pp. 361-376.

19 MERCADER UGUINA, Jesús R.: "El Reglamento de inteligencia artificial entra en la recta final, una primera lectura en clave laboral", *Revista General de Derecho del Trabajo y de la Seguridad Social*, (2024), 67, pp. 327-351.

Capítulo 2

LA INTELIGENCIA ARTIFICIAL Y EL DERECHO DEL TRABAJO EN CLAVE PERSPECTIVA DE GÉNERO

Mónica Ricou Casal

Profesora lectora. Estudios de Derecho y Ciencia política.
Universitat Oberta de Catalunya (UOC)

"Si cambias el modo en que miras las cosas, las cosas que miras cambian".
Wayne Dyer. Escritor norteamericano

RESUMEN: En este capítulo se pretende abordar los posibles efectos perniciosos de trasladar a la realidad virtual lo aún no resuelto en la realidad presencial. En efecto, hay sesgos en clave de perspectiva de género que presumiblemente van a seguir vivos si no se visibilizan y se aboga por poner límites.

En primer lugar, se hace eco de la normativa internacional, europea y nacional relativa a perspectiva de género, que es aplicable también cuando se hace uso de inteligencia artificial. Es menester ver también en el incipiente derecho digital los sesgos que claramente se visibilizan, con la consecuente involución de derechos.

En segundo lugar, se pondrá de relieve como el gran obstáculo de la mujer en el mercado laboral, es decir, la conciliación vida familiar y laboral, puede seguir siendo un obstáculo en los entornos virtuales. No ayuda que sean mayoritariamente hombres los que programan los modelos matemáticos, ya que se origina un problema de base de sesgo de género.

Como conclusión, para que no haya involución de derechos, se ve la necesidad de que en cada actuación que se realice se tenga en cuenta la perspectiva de género. Solo de esta forma, la IA representará a toda la población y no solo a la mitad de esta.

PALABRAS CLAVES: inteligencia artificial, límites, perspectiva de género, derechos, inversión

1. INTRODUCCIÓN

La inteligencia artificial (en adelante IA) representa un gran avance en la industria. Siguiendo a Ginés, se puede definir "*como aquella que aprende y desarrolla la mayoría de las tareas intelectuales de las personas humanas, incluido el desarrollo de inteligencia artificial.*"[1] En el marco de la evolución de la sociedad y del mercado de trabajo, es un salto cualitativo y cuantitativo de envergadura. Si hasta el momento la revolución industrial se había centrado en la sustitución de la fuerza física por máquinas, afectando a determinados grupos de categorías de personas trabajadoras, la nueva era iniciada pretende sustituir no ya la fuerza física sino la inteligencia humana. Pienso que complementar sería más adecuado que sustituir, ya que la mente humana es tan perfecta y sufre una evolución de niño a adulto que se hace difícil imaginar que esta inteligencia pueda ser sustituida.

Así las cosas, sí que sorprende y, por tanto, se debe alertar de que la normativa en perspectiva de género en la IA no se localice, no parece ser un valor o principio para tener en cuenta. Por supuesto que el Reglamento de IA europeo parte de los riesgos que conlleva la IA, y un sesgo es el género, pero no parece que haya medidas contundentes que resalten la perspectiva de género.

La incorporación de la mujer al mercado laboral no fue ajena de un traslado de los estereotipos de género al mismo. En suma, la distinción biológica de la mujer, siendo la que tiene la capacidad reproductora, le ha llevado y conlleva límites y obstáculos que aún hoy en día queda mucho por hacer.

Después de la experiencia y ahora que se está empezando a regular la aplicación de la IA y sus riesgos, ahora es el momento de que el uso creciente de aquella sea igualitario. Desde estas líneas no es otra la intención que alertar de ver cómo la conquista de derechos de las mujeres se pierde en el mundo virtual.

La normativa en perspectiva de género debe ser aplicable, no está excluida, es importante ver y detectar con rapidez y modificar los sesgos de género que se encuentren. Y crucial, un cambio tan trascendental no puede venir de la mano únicamente de un colectivo: las personas trabajadoras y los hombres, que por educación e historia, aun sin querer, tienden a reproducir patrones patriarcales.

2. LA INTELIGENCIA ARTIFICIAL NO ES AJENA A LA NORMATIVA EN PERSPECTIVA DE GÉNERO.

En este apartado se menciona la normativa en perspectiva de género a valorar en cada reglamentación que se lleve a cabo de la IA, no es un tema para nada baladí. De igual manera, se alerta de las brechas que ya se están observando en el derecho digital del derecho del trabajo, sin ánimo de exhaustividad por la extensión de la comunicación.

1 GINÉS I FABRELLAS, Anna: "Inteligencia artificial y sesgos" Blog NET21, julio 2024, p. 1

2.1. Normativa en perspectiva de género a aplicar.

En la realidad virtual, la normativa en perspectiva de género no queda excluida. Esto implica que siempre se debe tener presente desde las primeras normas, por esta razón vale la pena recordar su existencia. Así las cosas, no se van a cometer los mismos errores que inicialmente el ordenamiento jurídico cometió con la mujer, dificultándole, por ejemplo, su introducción al mercado laboral.

La igualdad de género fue incorporada al derecho internacional en la Declaración Universal de Derechos Humanos al adoptarse por la Asamblea General de las Naciones Unidas el 10 de diciembre de 1948[2]. En 1979, la Asamblea General aprobó la Convención sobre la Eliminación de todas las Formas de Discriminación contra la Mujer[3], una Carta Internacional de Derechos Humanos para las Mujeres. No se puede dejar de mencionar la Plataforma de acción de Beijing de 1995 y su seguimiento[4]. El 2 de julio de 2010, la Asamblea General de las Naciones Unidas vot**ó** por unanimidad la creación de un único organismo de la ONU para acelerar el progreso sobre igualdad de género y empoderamiento de la mujer: ONU Mujeres.

El empoderamiento de la mujer en el sector del Derecho digital es fundamental en el mundo del trabajo para hacer realidad la visión de la Agenda 2030 para el desarrollo sostenible de "*No dejar a nadie atrás"* y conseguir todos los objetivos de desarrollo sostenible[5]. En especial, se pretende la consecución del objetivo 5 "Lograr la igualdad de género y empoderar a todas las mujeres y niñas".

Para promover un mundo del trabajo justo, inclusivo y seguro, es esencial, también, el Plan de acción de la Organización Internacional del Trabajo (en adelante OIT) sobre Igualdad de género 2022-2025, que se vincula con los objetivos de desarrollo sostenible de la Agenda 2030[6]. En el contexto de la OIT es crucial para poder llegar a conseguir el reto del trabajo decente. En la normativa de la Unión Europea es importante mencionar la Estrategia

2 NU. Declaración Universal de Derechos Humanos. Adoptada y proclamada por la Asamblea General en su resolución 217 A(III), de 10 de diciembre de 1948. Disponible en: https://www.ohchr.org/sites/default/files/UDHR/Documents/UDHR_Translations/spn.pdf

3 NU. Convención sobre la eliminación de todas las formas de discriminación contra la mujer. Aprobado el 18 de diciembre de 1979. En vigor el 3 de septiembre de 1981, disponible en: https://www.ohchr.org/es/instruments-mechanisms/instruments/convention-elimination-all-forms-discrimination-against-women

4 ONU Mujeres. Declaración y Plataforma de Acción de Beijing. 1995, disponible en: https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/CSW/BPA_S_Final_WEB.pdf

5 NU. Agenda 2030 para el desarrollo sostenible. El 25 de septiembre de 2015, la Asamblea General de Naciones Unidas aprobó por unanimidad la Agenda 2030. Disponible en: https://www.pactomundial.org/que-puedes-hacer-tu/ods/

6 OIT. "Plan de Acción de la OIT sobre Igualdad de Género 2022-2025" Ginebra 2022. Disponible en: https://researchrepository.ilo.org/esploro/outputs/report/Plan-de-acci%C3%B3n-de-la-OIT/995264839902676_Vid: Estrategia para la igualdad de género - Comisión Europea (europa.eu)

para la Igualdad de género 2020-2025[7] para atajar las brechas en el mercado de trabajo (no solo en el trabajo tradicional).

En el ordenamiento jurídico español, la igualdad es un pilar fundamental, así lo prescribe tanto en sus artículos 1, 9.2 y art. 14 de la Constitución española[8]. Se debe tomar como referencia tanto la LO 3/2007, de 22 de marzo, para la igualdad efectiva entre mujeres y hombres,[9] y la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación[10].

2.2. En la incipiente normativa en IA, ¿se atisban sesgos de género en materia de derecho laboral?

Como normativa puntera en el ámbito de la inteligencia artificial se debe destacar el reciente y esperado Reglamento UE 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024[11] (en adelante Reglamento IA). Al tratarse de un Reglamento, no se olvide que es de aplicación directa en todos los Estados miembros, es decir, no requiere trasposición. La clave es que regula normas armonizadas que se basan en dos elementos: sus riesgos potenciales y su nivel de impacto.

Así las cosas, el Reglamento IA tiene como objetivo al ser humano, garantizando un óptimo nivel de protección de la salud, la seguridad, y los derechos fundamentales previstos en la Carta Europea de Derechos Fundamentales de la Unión Europea[12] (en adelante, CEDF). El sistema de IA se define: "*como un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos y virtuales*" art. 3.1 del Reglamento IA.

---

7 UE. Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones Una Unión de la igualdad. "Estrategia para la Igualdad de género 2020-2025" de 5 de marzo de 2020. Disponible en: EUR-Lex - 52020DC0152 - EN - EUR-Lex (europa.eu)

8 Constitución española de 27 de diciembre de 1978. Entrada en vigor el 29 de diciembre (BOE 29 de diciembre).

9 LO 3/2007, de 22 de marzo, para la igualdad efectiva entre mujeres y hombres. BOE núm. 71, de 23 de marzo. Disponible en https://www.boe.es/buscar/act.php?id=BOE-A-2007-6115

10 Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación (BOE de 13 de julio). Disponible en https://www.boe.es/buscar/act.php?id=BOE-A-2022-11589

11 Reglamento UE 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (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 de 12 de julio de 2024)) disponible en: Reglamento - UE - 2024/1689 - EN - EUR-Lex (europa.eu). Entró en vigor a los 20 días de su publicación en el DOUE y será aplicable (salvo determinadas disposiciones) a partir de agosto de 2026.

12 Carta de los derechos Fundamentales de la Unión Europea (2016/C 2022/02) DOUE de 7 de junio de 2016 Disponible en EUR-Lex - 12016P/TXT - EN - EUR-Lex (europa.eu)

El Reglamento IA se basa en un enfoque centrado en el riesgo, de esta forma clasifica los sistemas de IA en función del riesgo: riesgo inaceptable, alto riesgo, riesgo limitado y riesgo nulo. De alto riesgo son los sistemas de IA de uso en la rama del derecho del trabajo y seguridad social. El motivo es que implican riesgos altos en la salud, seguridad y derechos fundamentales de las personas, por ejemplo, evaluar la conducta de las personas trabajadoras[13].

Como primer apunte a resaltar en materia de derecho del trabajo y seguridad social es que no se podría entender en el marco normativo europeo el Reglamento de IA sin la Directiva relativa a la mejora de las condiciones laborales en el trabajo en plataformas digitales (en adelante Directiva de Plataformas)[14], deben ir unidas y de la mano. También es menester poner de relieve el Reglamento General de Protección de datos[15]

La Directiva de Plataformas, es una novedad de calado dentro de la rama incipiente del Desarrollo Algorítmico del Trabajo. El objetivo es que, al igual que la Ley Rider en el ordenamiento jurídico español,[16] se pretende garantizar que se cumplan los derechos de las personas trabajadoras en plataformas, regular el uso de algoritmos en el mercado de trabajo y atajar el abuso de la figura del falso autónomo. Es una buena oportunidad para el ordenamiento jurídico español de revisar la Ley Rider.

Las plataformas digitales cada vez tienen más peso en nuestra vida diaria, acelerándose su uso desde la pandemia de la COVID-19, a causa del confinamiento y del teletrabajo. Se podrían diferenciar entre plataformas de trabajo *on line* y plataformas de trabajo i*n situ*[17]. El sesgo principal de trabajo en estas plataformas es que se repiten los roles de género del patriarcado, es decir, en similares términos que ocurrió en los inicios de la introducción de la mujer al mercado de trabajo.

13 GINÉS I FABRELLAS, Anna: "Inteligencia artificial y sesgos" Blog NET21, julio 2024,p. 4

14 Directiva relativa a la mejora de las condiciones laborales en el trabajo en plataformas digitales. En fecha 11 de marzo de 2024, el Consejo de la Unión Europea confirmó el acuerdo provisional sobre la Directiva relativa al trabajo en plataformas digitales que se había alcanzado el 8 de febrero de 2024, la Presidencia del Consejo y los negociadores del Parlamento Europeo. En fecha 23 de octubre de 2024 se publica la Directiva (UE) 2024/2831 del Parlamento Europeo y del Consejo relativa a la mejora de las condiciones laborales en el trabajo en plataformas. En vigor el 1 de diciembre. Disponible en https://eur-lex.europa.eu/legal-content/es/ALL/?uri=CELEX:32024L2831

15 Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales. (DOUE de 4 de mayo de 2016) Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A32016R0679

16 Ley 12/2021, de 28 de septiembre, por la que se modifica el texto refundido de la Ley del Estatuto de los Trabajadores, aprobado por el RD Legislativo 2/2015, de 23 de octubre, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales. (BOE de 19 de septiembre) Disponible en https://www.boe.es/buscar/act.php?id=BOE-A-2021-15767

17 "*En las plataformas de trabajo online, las personas trabajadoras llevan a cabo a través de internet los trabajados asignados, por ejemplo, servicios de traducción. En las plataformas de trabajo in situ, las personas trabajadoras realizan las tareas en persona, en ubicaciones físicas concretas, como por ejemplo un servicio de taxi o la prestación de cuidados*" Ministerio de Asuntos Económicos y Transformación Digital. *Perspectiva global sobre las mujeres, el trabajo y las plataformas digitales de trabajo* Digital Future Society 2022, pág. 15

En estos términos, para un 40% de las mujeres el trabajo en plataformas representa su principal fuente de ingresos. En los países en vía de desarrollo, el trabajo de plataformas es la principal fuente de ingresos para más de la mitad de las mujeres que trabajan en las mismas. En las economías avanzadas esa proporción se da en un tercio de las trabajadoras[18].

La buena intención de la legislación reguladora de las plataformas no se pone en duda, a pesar de todo, es aún incipiente y se requiere una gran labor de los operadores jurídicos, incluida la inspección de trabajo. Son relaciones que se enmascaran bajo una relación por cuenta propia cuando cumplen todos los requisitos para ser consideradas relaciones trabajadoras por cuenta ajena.

Trabajar en plataformas puede implicar que la oferta de trabajo sea a nivel internacional, lo que genera una intensa competencia internacional donde la educación de que el hombre lo puede hacer mejor sigue presente. "*Un cliente me dio a entender que la tarea podía ser demasiado complicada para mí porque soy mujer*." (trabajadora de Upwork, encuestada, Sudáfrica)[19].

El sesgo de género de mayor importancia para trabajar para las mujeres es la conciliación de la vida familiar y laboral. A muchas trabajadoras no les queda otra opción que conectarse por las noches y así se aseguran de recibir trabajo regular, quedando, por otro lado, invisibilizadas. En efecto, el 78% de las mujeres que usan plataformas de micro tareas desarrolla su trabajo por las noches. Este porcentaje se incrementa en los países en desarrollo, al 85 %.

Así las cosas, incluso mujeres con niveles de estudio alto, la decisión de poderse quedar teletrabajando pesa en la elección de una oferta de trabajo. De hecho, las personas trabajadoras en plataformas *on line* están altamente cualificadas. Sin duda, se limita la posibilidad de acceder a empleos estables, siendo las plataformas las que ofrecen flexibilidad para integrarse en el mercado de trabajo. Se puede derivar que el trabajo en plataformas digitales reafirma los roles de las mujeres en la reproducción social y las aleja de las relaciones sociales y profesionales, de los contactos, de la presencia física que se puedan dar en los lugares de trabajo o incluso en cursos de formación presenciales[20].

Tampoco se ha eliminado la brecha de género. En el trabajo en plataformas *on line* hay más hombres que mujeres; estas representan aproximadamente un 40%. Hay un país, Estados Unidos, donde sí, el porcentaje de mujeres llega al 50 %[21]. También se puede hablar de segregación de género, con base en que en las plataformas de trabajo autónomo los hombres son los que realizan la mayoría de las tareas que se relacionan con la tecnología, diseño creativo y multimedia. Las mujeres se dedican a trabajos como los servicios profesionales, ámbito

18 Ministerio de Asuntos Económicos y Transformación Digital. *Perspectiva global sobre las mujeres, el trabajo y las plataformas digitales de trabajo* Digital Future Society 2022, p. 16,

19 ídem

20 Ministerio de Asuntos Económicos y Transformación Digital. *Perspectiva global sobre las mujeres, el trabajo y las plataformas digitales de trabajo* Digital Future Society 2022, pág. 17

21 Ibídem p. 19,

jurídico, traducción, edición o redacción, o incluso marketing[22]. Se reitera también que las brechas salariales de género varían en función del trabajo, del tipo de tarea, según plataforma y si se trata de un país en vías de desarrollo o desarrollado[23].

Como señala Ginés, en general en los sistemas de IA es verdaderamente preocupante que abunden los sesgos y estereotipos de género, principalmente porque son herramientas pensadas para ganar en productividad en el día a día[24]. De esta forma en un estudio se han analizado más de 5000 imágenes generadas por *Stable difusión* y la conclusión resultó que se reproducían estereotipos de género, de raza o de clase de la realidad no virtual, e incluso en mayor medida, por estar entrenados los modelos de IA generativa con datos de internet que ya incluyen sesgos[25].

3. LA INTELIGENCIA ARTIFICIAL EN EL MERCADO DE TRABAJO NO PUEDE SER CREADA POR Y PARA LOS HOMBRES

En las nuevas formas de trabajo emergentes, cada vez implantadas con más fuerza, quien las crea debe poder representar a toda la población. En otros términos, los hombres son los que reciben la formación en ingeniería e informática en mucha mayor proporción que las mujeres. En cambio, la proporción de mujeres que reciben educación en medicina y ciencias naturales es mayor.

Se debe poner en alerta que el desarrollo de los algoritmos de las plataformas de trabajo *on line* se diseña por hombres y es volver a revivir el sesgo estructural existente desde que se inicia la rama laboral del derecho. Hay precedentes suficientes para no reproducir lo mismo, lo que implica que la mujer debe seguir luchando contra una normativa y educación patriarcal.

Es necesario un diálogo social global para abordar este reto de la igualdad en todo lo que representa el derecho digital del trabajo y el impacto de la IA Generativa en el mercado laboral. Asimismo, la OIT tiene un papel más que destacado. En este sentido, el Informe de la OIT de 21 de agosto de 2023 señala que el trabajo administrativo, realizado mayoritariamente por mujeres, sería la categoría con mayor exposición tecnológica, con una cuarta parte de las tareas consideradas altamente expuestas y más de la mitad de las funciones con una exposición de nivel medio. En otros grupos profesionales, como directivos, profesionales y técnicos, es un porcentaje pequeño de las tareas que se consideran más expuestas[26].

22 Ibídem p. 19,

23 Ibídem p. 20,

24 GINÈS I FABRELLAS, Anna: "Inteligencia artificial y sesgos" Blog NET21, julio 2024, p. 1

25 Ídem.

26 OIT. Generative AI and Jobs. *A global analysis of potencial effects on job quantity and quality* Informe OIT. 21 de agosto de 2023. Disponible en: https://www.ilo.org/publications/generative-ai-and-jobs-global-analysis-potential-effects-job-quantity-and

La OIT también pone de manifiesto que los efectos de la inteligencia artificial generativa van a tener efectos diferentes entre hombres y mujeres. El motivo es la sobrerrepresentación de las mujeres en el trabajo administrativo. Sea como fuere, como los trabajos administrativos han estado tradicionalmente feminizados sobre todo cuando los países se desarrollaban económicamente, indica el informe de la OIT que uno de los resultados de la nueva tecnología puede ser que según qué trabajos administrativos nunca surgirán en los países de renta baja.

La OIT pone el énfasis en la necesidad del diseño de políticas para una transición justa e igualitaria. "*La voz de las personas trabajadoras, la capacitación y una protección social adecuada serán claves para gestionar la transición."* De lo contrario, se corre el riesgo de que solo unos pocos países y participantes en el mercado bien preparados se beneficien de la nueva tecnología.[27]

El trabajo digital que se basa en las competencias TIC conlleva para las mujeres una brecha por sus carencias formativas en estas áreas. Esto conlleva que se les impide el acceso a determinados puestos de trabajo donde estas competencias son cada vez más demandadas. La igualdad en derechos digitales es urgente para un acceso igualitario al empleo y que no se repitan roles patriarcales. En suma, la creación de algoritmos y la inteligencia artificial debe ser creada por mujeres y hombres[28].

En el sistema de Indicadores de Género y TIC (SIGTIC) se calcula el grado de igualdad de género siguiendo una serie de indicadores: frecuencia de uso de ordenador e internet, del correo electrónico, búsqueda de empleo, ocio, educación, salud, acceso a la administración pública, entidades bancarias vía internet, etc. Esta brecha digital encuentra su base principalmente en la educación familiar y en la educación escolar, donde se promueve que la investigación científica es más propia de niños que de niñas. Se seguiría el prototipo de mujer sumisa[29].

Las niñas no tienen motivación para dedicarse a las carreras STEM (ciencia, tecnología, ingeniería y matemáticas), y provoca falta de representación en estos campos. Así las cosas, el 17 % de estudiantes de informática son mujeres. Además, no se puede dejar de poner de relieve el sector de los juguetes y videojuegos y su influencia para perpetuar roles de género. Es importante subrayar la carencia de referentes femeninos en las carreras STEM. En este sentido, solo el 0.5% de las adolescentes de 15 años de los países de la OCDE en 2018 ponía el foco en su carrera profesional en el ámbito de las nuevas tecnologías. Promover políticas de captación del interés de las mujeres en las TIC, es prioridad de primer orden[30].

El colectivo femenino representa el 20% de las personas trabajadoras con roles técnicos, el 12% de las investigadoras y el 6% de las desarrolladoras de software profesional. De

27 *Ídem*

28 PÉREZ LÓPEZ, José Iván: "Brecha digital, género y derechos laborales", ADAPT University Press, *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo,* Volumen 11, número 2, abril-junio 2023, p. 362

29 Ibídem, p. 365

30 Ibídem, p. 367

estos datos se deriva la carencia de perspectiva de género en la industria de la IA, en los datos de entrenamiento que con posterioridad se trasladarán al contenido generado por el sistema[31].

En fin, visto que no hay previsiones normativas, hay que tomar en consideración propuestas como la de CC.OO., que pretende incluir en los convenios colectivos negociados por este sindicato la siguiente cláusula: "*toda persona que programe o adquiera algoritmos de gestión deba ser formada por la empresa para conocer debidamente los riesgos de parcialidad y discriminación y adoptar todas las medidas posibles para reducirlos*"[32].

4. CONCLUSIONES

Con base en lo analizado en los anteriores párrafos, urge, debido al imparable crecimiento de la IA que va a ser fundamental tanto en la vida personal, laboral y académica, sentar las bases de:

En primer lugar, la educación a todos los niveles no puede reproducir estereotipos de género que claramente representarán una involución de derechos para el colectivo de mujeres trabajadoras.

En segundo lugar, los métodos de entrenamiento de la IA y los algoritmos han de tener la supervisión suficiente para que no sean generadores de discriminaciones tanto de género como de cualquier otro tipo. La formación para que en las decisiones de creación de IA haya paridad entre hombres y mujeres sería la clave para acercarnos más al objetivo de igualdad.

En tercer lugar, en el ordenamiento jurídico español la IA, es decir, el derecho digital laboral, se debe ver reflejado en el Estatuto de los Trabajadores, para que la normativa se adecue a la realidad y no se quede atrás. Los convenios colectivos deberían tener como contenido obligatorio estas materias para poder llegar a trabajadores y empresarios, teniendo como faro que la digitalización llegue a todos y todas.

Si como afirma BELTRÁN DE HEREDIA "*Blindar a la mente inconsciente frente a todas las intromisiones externas es el mejor legado que podríamos dejar a las generaciones futuras*"[33], una de esas intromisiones es el androcentrismo en el mercado de trabajo, en el derecho laboral, que en muchas ocasiones ha invisibilizado o no valorado el talento femenino.

31 GINÈS I FABRELLAS, Anna: "Inteligencia artificial y sesgos" Blog NET21, julio de 2024, p. 2

32 RODRÍGUEZ FERNÁNDEZ, María Lu:: "Inteligencia artificial, género y trabajo", Edit. *Junta de Andalucía. Temas Laborales. Revista andaluza de trabajo y bienestar social*. Número 171/2024

33 BELTRÁN DE HEREDIA, Ignasi: "*Hacía el estatuto del yo inconsciente de la persona trabajadora*", *Editorial La Ley, Trabajo y Derech*o, N.º 19, junio 2024, p. 14

5. BIBLIOGRAFÍA

BELTRÁN DE HEREDIA, Ignasi: "*Hacía el estatuto del yo inconsciente de la persona trabajadora*", *Editorial La Ley, Trabajo y Derecho*, nº 19.junio 2024.

GINÉS I FABRELLAS, ANNA: "Inteligencia artificial y sesgos" *Blog NET21,* julio 2024.

PÉREZ LÓPEZ, José Iván: "Brecha digital, género y derechos laborales", ADAPT University Press, *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo,* Volumen 11, número 2, abril-junio 2023.

RODRÍGUEZ FERNÁNDEZ, María Luz: "Inteligencia artificial, género y trabajo", Edit. *Junta de Andalucía. Temas Laborales. Revista andaluza de trabajo y bienestar social*. núm. 171/2024

Capítulo 3

THE INTERPLAY BETWEEN AI ACT AND GDPR

Nidal Askar

Ph.D. Candidate in Law
The University of Strasbourg, the Centre for European and International Studies

ABSTRACT: This analysis explores the interaction between the General Data Protection Regulation (GDPR) and the EU Artificial Intelligence Act (AI Act), examining their regulatory frameworks and roles as of July 2022. While the GDPR focuses on fundamental rights protection and governance of personal data processing, the AI Act aims at ensuring the safe development and deployment of AI systems, particularly in high-risk scenarios. This article scrutinizes the interplay between the two regulations, examining how they complement each other in addressing societal risks associated with AI. Although the GDPR and the AI Act generally do not conflict, areas of tension particularly regarding specific legal provisions, concrete processing operations and their purpose and means. It remains to be seen whether these regulations can effectively complement each other.

KEYWORDS: Artificial Intelligence, AI Act, GDPR, Data Protection, European Union

1. INTRODUCTION

Although the GDPR and the EU AI Act appear to be similar legal instruments, as both include regimes of accountability, governance, and oversight, they serve different purposes.[1]

The EU's data protection law is predominantly defined by the GDPR. The current framework of the GDPR is based on the 1995 Data Protection Directive, the right to data protection in Article 8 of the Charter of Fundamental Rights (CFR), the right to privacy in Article 7 of the CFR, and the primary legal basis in Article 16 TFEU.

The EU AI Act is predominantly a product safety law aimed at ensuring the safe technical development and deployment of AI systems; and apart from a few exceptions, it does not confer any rights to individuals.[2] It adopts a horizontal approach, unlike other product safety legislations, as it is not specific to any particular sector but is instead dedicated to the use of AI systems.[3] The EU AI Act is based on a risk-oriented approach, categorizing AI systems into four levels of risk: unacceptable, high, limited, and minimal/no-risk. AI systems posing an "unacceptable risk" will be prohibited, while "high-risk" AI systems will be subject to stringent obligations before they can be introduced to the EU market.[4] Unlike the GDPR, the EU AI Act implements a comprehensive risk categorization and imposes distinct obligations based on these categories. Most requirements under the EU AI Act are applicable solely to high-risk AI systems (as outlined in Article 6 and Annex III of the EU AI Act). Specific obligations also apply to various AI systems, such as general-purpose AI models, along with transparency requirements for systems like emotional categorization systems.[5] Most of the provisions of the AI Act focus on high-risk systems, outlining obligations for providers, users, and other participants throughout the AI value chain, and establishing conformity assessment procedures to be followed for high-risk AI systems.[6]In contrast, the GDPR is a fundamental rights legislation that grants individuals extensive rights regarding the processing of their personal data. The GDPR's subject matter and scope encompass the processing of personal data with the aim of safeguarding the fundamental rights and freedoms of individuals, as outlined in Article 1 of the GDPR.[7] The definition of 'processing' personal

1 CEDPO AI Working Group, *AI and Personal Data A Guide for DPOs "Frequently Asked Questions"* (2023) p. 10.

2 M. VEALE and F. Z. BORGESIUS, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach' (2021) 22 *Computer Law Review International* 97–112 at 97.

3 M. EBERS, 'Standardizing AI - The Case of the European Commission's Proposal for an Artificial Intelligence Act' (2021) p. 13.

4 'Ethics guidelines for trustworthy AI | Shaping Europe's digital future' (April 2019).

5 J. CLARK, M. DEMICRAN, K. KETTAS, 'Europe: The EU AI Act's relationship with data protection law: key takeaways' (April 2024).

6 M. EBERS, 'Standardizing AI - The Case of the European Commission's Proposal for an Artificial Intelligence Act', p. 1.

7 Article 1 of the *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)* (2016).

data is extensive, encompassing almost any interaction with personal data. Consequently, any stage in the lifecycle of an AI model involving personal data may fall under the GDPR's jurisdiction.[8] At first glance, this approach may seem quite straightforward; however, there are important differences to consider.[9] The AI Act can initially be viewed as a reinforcement or complement to the GDPR, given that certain principles of the GDPR are reflected in certain provisions of the AI Act.[10] These shared principles can be found in provisions relating to "risk management systems[11]", "data governance[12]", "technical documentation[13]", "human supervision[14]", and "transparency[15]." In other terms, the AI Act commits to establishing a detailed and technologically tailored framework designed to complement the GDPR in fostering responsible AI innovation within Europe, while also upholding the fundamental rights and values of the Union.[16]

8 CEDPO AI Working Group, *AI and Personal Data A Guide for DPOs "Frequently Asked Questions"*, p. 10.

9 CEDPO AI Working Group, AI and Personal Data A Guide for DPOs "Frequently Asked Questions", p. 10.

10 H. Avocats, 'Actualités juridiques et digitales'.

11 Article 9 of the Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) (2024); Article 35 – Data Protection Impact Assessment (DPIA) of the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

12 Article 10 of the Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance).

13 Article 11 of the Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle et modifiant les règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 et (UE) 2019/2144 et les directives 2014/90/UE, (UE) 2016/797 et (UE) 2020/1828 (règlement sur l'intelligence artificielle) (Texte présentant de l'intérêt pour l'EEE) (2024).

14 Article 14 of the Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance).

15 Article 13 (Transparency and provision of information to deployers) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance); Article 5(1)(a) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

16 CEDPO AI Working Group, AI and Personal Data A Guide for DPOs "Frequently Asked Questions", p. 10.

Consequently, organisations need to carefully map the regimes of the GDPR and the EU AI Act carefully to determine which bodies are subject to the requirements of the AI Act, the GDPR, or both.[17] In some points, there are tensions and overlaps between the GDPR and AI. In this respect, the GDPR could be perceived as both insufficient and a hindrance to technological advancement.

The legal analysis begins with the overlapping statutory objectives and scopes of application of the GDPR and the EU AI Act. Following this, the principal divergences of the chosen statutory notions of both legal instruments are outlined. In the following section, to delve into the tensions between the two legal acts, specific points will be addressed. Finally, it will become clear that for the purpose of proper interpretation and application of both AI regulation and the GDPR, procedural safeguards or concrete key provisions to reconcile conflicting legal norms will be necessary.

2. TENSIONS AND OVERLAPPING POINTS IN THE GDPR AND THE EU AI ACT

2.1. Areas of Tensions in the Application

While some view the GDPR as a robust mechanism for safeguarding personal data, others argue the regulation inadequate for governing the processing of personal data by AI systems.[18] For instance, critics highlight that the GDPR's requirements may not fully address the complexities of data processing in AI training processes. This divergence in perspectives highlights the tensions as well between the GDPR and AI Act, particularly in scenarios where AI systems process personal data. Such tensions create complexities that necessitate careful navigation of both legal frameworks.

Since the AI Act and EU data protection law are enforced concurrently, the European Data Protection Board (EDPB) and the European Data Protection Supervisor (EDPS) have emphasized during negotiations the necessity of clearly avoiding any inconsistencies or potential conflicts between the AI Act and data protection regulations.[19] However, these concerns have largely been overlooked.[20] In some respects, there is a conflict between the GDPR and AI. There are specific concerns the alignment of AI technologies with core principles of data protection law such as purpose limitation, data minimisation, transparency and lawfulness.[21]

17 'Key Issue 6: Interplay with GDPR - EU AI Act'.

18 N. WALLACE and D. CASTRO, 'The Impact of the EU's New Data Protection Regulation on AI' (March 2018).

19 'EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) | European Data Protection Board' para.57.

20 M. EBERS, 'Truly Risk-Based Regulation of Artificial Intelligence - How to Implement the EU's AI Act' (2024) p. 19.

21 C. KUNER, F. H. CATE, O. LYNSKEY, C. MILLARD, N. NI LOIDEAIN, and D. J. B. SVANTESSON, 'Expanding the artificial intelligence-data protection debate' (2018) 8 *International Data Privacy Law* 289–92 at 289–90.

AI typically involves the collection of vast amounts of data, particularly during the training phase, and many AI systems possess a broad potential range of applications, such as mimicking human-like intelligence, which makes the clear definition of "processing purposes" challenging.[22] Consequently, the issue arises regarding the quantity and type of data required to ensure successful training.[23] The connection to the purpose limitation within the data minimisation principle serves as an entry point for AI-specific considerations, which emphasize the system's overall functionality rather than focusing solely on individual processing operations.[24] This aligns with the accuracy requirement under the AI Act(Article 15), which seeks to mitigate risks to the health, safety, and fundamental rights of individuals. Furthermore, it supports the overarching aim of data protection law, specifically in safeguarding fundamental rights and freedoms, including the right to non-discrimination. Nonetheless, at the same time it poses a challenge to the objective of limiting the quantity and the intrusiveness of personal data to safeguard the right to data protection.[25] Legal requirements concerning the quantity and quality of training datasets necessary for successful training are outlined in Article 10 of the AI Act. If the processing is intended for training a high-risk AI system, the principles of adequacy, relevance, and necessity outlined in Article 5(1)(c) of the GDPR, which are purpose-related, must be construed and implemented in accordance with Articles 15 and 10 of the AI Act.[26] Therefore, the determination of the appropriate quantity and type of data that are adequate, relevant, and necessary for the training of a specific AI system must be evaluated based on the criteria of relevance, representativeness, error-free status, and completeness as stipulated in Article 10(3) of the AI Act, and to ensure accuracy as required by Article 15(1) of the AI Act.[27] As a result both Article 5(1)(c) of the GDPR and Article 10 and 15 of the AI Act must be construed and applied in concordance, if any tension arises a balance must be struck.[28]

The AI Act risks establishing parallel enforcement structures alongside data protection authorities, potentially resulting in legal uncertainty.[29] Although the EDPB and the EDPS have emphasized that national data protection authorities should be tasked with enforcing

22 KETTAS, 'Europe'.

23 M. WINAU, 'Areas of Tension in the Application of AI and Data Protection Law' (2023) 9 *European Data Protection Law Review* 123–35 at 132.

24 M. WINAU, 'Areas of Tension in the Application of AI and Data Protection Law', 132.

25 Ibid.

26 Ibid.

27 Ibid.

28 Ibid.

29 P. HAJDUK, AI Act and GDPR: On the path towards overlap of the enforcement structures, October 2023, RAILS Blog, https://blog.ai-laws.org/

the AI Act[30], Article 70 of the AI Act delegates the authority to designate competent bodies to the Member States.[31] This is likely to result in the authorisation of various entities with overlapping competencies, as seen in the case of the recently established Spanish Agency for the Supervision of AI.[32]

On the other hand, there is an evident overlap between several data protection principles and requirements outlined in the EU AI Act to ensure the secure development and deployment of AI systems. For instance, while both the GDPR and the AI Act impose transparency, they do so with different scopes and requirements.[33] The interplay between AI and data protection is explicitly acknowledged in the EU AI Act, which affirms that it does not undermine the GDPR.[34] The legal basis of the EU AI Act is Article 16 of the Treaty on the Functioning of the European Union ("TFEU"), which mandates the EU to establish regulations concerning the protection of individuals regarding the processing of personal data.[35]

2.2. The Overlapping Statutory Objectives and Scope of Application of The Two Regulations:

The subject matter of the GDPR, as outlined in Article 1, paragraph 1, aims to ensure the free movement of personal data while safeguarding the fundamental rights and freedoms of natural persons from the risks associated with personal data processing. Similarly, Recital 1 of the EU AI Act aims to establish a uniform legal framework for the development, marketing, and use of AI systems, thereby ensuring the free movement of AI-based goods and services and

30 'EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) | European Data Protection Board' para 48.

31 See also Recital 157 AI Act: "This Regulation is without prejudice to the competences, tasks, powers and independence of relevant national public authorities or bodies which supervise the application of Union law protecting fundamental rights, including equality bodies and data protection authorities."

32 P. SOLIMANO, 'Spain Just Created the First European AI Supervision Agency' (August 2023) , https://decrypt.co/153482/.

33 The GDPR enshrines the principle of transparency to support the exercise of data subjects' rights, such as the rights to erasure, rectification, and data portability, as outlined in Articles 15-22. On the other hand, the AI Act imposes transparency obligations specifically for high-risk AI systems (Article 13) and certain other AI systems (Article 50). Additionally, Article 13 of the AI Act prioritizes the interests of the AI system deployer rather than those of the end user or data subject.

34 Recital 9, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(Ordinary legislative procedure: first reading).References to articles and recitals introduced or changed by the Parliament will be labelled in this paper as "the EU AI Act".

35 Article 16 of the Treaty on the Functioning of the European Union, which is related to the protection of personal data

protecting public interests, including the explicit protection of individual fundamental rights. Consequently, both regulations strive to achieve a balance between ensuring a free market and protecting economic and public interests in the processing of personal data and the use of AI, while also safeguarding Union values, fundamental rights, and principles that may be endangered by these activities, which demonstrates parallelism in their regulatory purposes.

However, the AI Act and the GDPR differ essentially in their scope of application. According to Article 2 of the AI Act, it applies to providers, users, and other actors in the AI value chain, including importers and distributors of AI systems used in or placed on the EU market, regardless of their location, whereas referring to Articles 2 and 3 of the GDPR, it applies to controllers and processors of personal data within the EU, or to entities offering goods or services to, or monitoring the behavior of, individuals within the EU. Moreover, the GDPR, does not address AI as a specific method of data processing, since there are no special AI provisions in the GDPR apart from Article 22.[36] Besides, not all AI or machine learning systems need to process personal data; therefore, in those cases, the GDPR will not be applicable.[37]

Therefore, AI systems that do not process personal data or handle the personal data of non-EU individuals may be subject to the AI Act but not the GDPR.[38] For example, a company developing an AI system that analyses non-personal data for market trends would fall under the AI Act but would not be governed by the GDPR.

The GDPR applies to AI systems, insofar as personal data is involved at any stage of an AI system's lifecycle. Under Article 5(1)(a) of the GDPR, "processing" does not denote a specific technical processing method.[39] Instead, it encompasses a wide range of automated (algorithm-based) processing techniques that produce outputs aligned with predetermined human-defined objectives.[40] Compared to the GDPR, the AI Act has a broader scope in terms of application to entities, yet is more limited regarding the methods and purposes of processing.[41] As a result, there is an overlap of matters when personal data is processed within an AI system.[42] In the light of this, it is clear that the regulatory objectives of both legal instruments share significant similarities.

36 I. SPIECKER and G. DÖHMANN, 'AI and Data Protection' in L. A. DIMATTEO, C. PONCIBÒ, M. CANNARSA (eds.), The Cambridge Handbook of Artificial Intelligence, (Cambridge University Press, 2022), pp. 132–45 p. 133.

37 CEDPO AI Working Group, *AI and Personal Data A Guide for DPOs "Frequently Asked Questions"*, p. 10.

38 'International: The interplay between the AI Act and the GDPR - AI series part 1' (November 2023).

39 M. WINAU, 'Areas of Tension in the Application of AI and Data Protection Law', 124.

40 For more detail and critical regarding the definition of Article 5(1)(a) of the AI Act M. EBERS, V. R. S. HOCH, F. ROSENKRANZ, H. RUSCHEMEIER, and B. STEINRÖTTER, 'The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)' (2021) 4 *J* 589–603 at 589–90.

41 M. WINAU, 'Areas of Tension in the Application of AI and Data Protection Law', 124.

42 Ibid.

Article 2(7) of the AI Act states that the regulation does not affect EU Data Protection Law.[43] Hence, AI systems based on personal data are required to comply with both the AI Act and GDPR. Moreover, the AI Act does not include any explicit requirements for AI systems to comply with GDPR provisions to be placed on the EU market, despite recommendations from both the EDPS and the EDPB that co-legislators to include such a requirement in the AI Act. Specifically, both authorities advised the certification of high-risk AI systems should explicitly verify compliance with the GDPR. [44] Conversely, Recital 63 of the AI Act states that "*an AI system is classified as a high-risk AI system should not be interpreted as indicating that the use of the system is lawful under other acts of Union law or under national law compatible with Union law, such as on the protection of personal data, on the use of polygraphs and similar tools or other systems to detect the emotional state of natural persons*." It continues to clarify that "*the AI Act should not be understood as providing for the legal ground for processing of personal data, including special categories of personal data, where relevant, unless it is specifically provided for otherwise in the AI Act.*"[45] Since the AI Act and the GDPR apply simultaneously, both the EDPB and the EDPS stressed that it is vital to clearly avoid any inconsistencies and possible conflicts with the GDPR in the AI Act. [46] Veritably, certain concepts and provisions of the AI Act overlap with EU data protection law, which may result in legal ambiguities and varying interpretations.

Within a practical context, the AI Act provides a double layered protection, the first phase covers the development or production of AI systems, and the second phase pertains to their utilization.[47] When a high-risk AI system is developed, the obligations arising from Chapter III, particularly Article 16, apply. If personal data is processed within this high-risk AI system, the GDPR will also apply to the provider as the controller.[48]

Non-personal data can turn into personal data throughout its lifecycle if additional information that can be used to identify a person becomes available.[49] Moreover, for the learning

43 According to Art 2(7) of the AI Act, data processing for the purpose of ensuring bias detection in high-risk AI systems in its Article 10(5) and data processing in AI regulatory sandboxes in its Article 59 are explicitly excluded.

44 'EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) | European Data Protection Board' para 23; 'EDPS Opinion 44/2023 on the Proposal for Artificial Intelligence Act in the light of legislative developments | European Data Protection Supervisor' (September 2024) para 27.

45 Recital 63, *European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(Ordinary legislative procedure: first reading).References to articles and recitals introduced or changed by the Parliament will be labelled in this paper as "the EU AI Act".*

46 para 57 'EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) | European Data Protection Board'.

47 M. JACOBS and J. SIMON, 'Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed By the European Commission' (2022) 1 *Digital Society* 6.

48 M. WINAU, 'Areas of Tension in the Application of AI and Data Protection Law', 125.

49 Ibid.

of AI systems, data is processed on two levels: one for training and the other for validation purposes.[50] A small amount of personal data that is not eliminated from anonymized processed data can trigger the application of the GDPR at one of these levels.[51]

As a result, it can be argued that the interpretation and application of these overlapping areas must be done in coordination with each other for the sake of an effective application of both regulations, which the AI Act does not address it.[52]

2.2.1. Overlapping Principles in the Both Regulations

Both the GDPR and the AI Act impose **transparency** obligations; however, the scope and requirements differ between the two regulations.[53] Article 12 of the GDPR is important for the governance of machine learning systems, as in many cases, deployers are not aware of their data being processed.[54] Article 50(4) of the EU AI Act contributes to this general requirement by imposing obligations for the creation of deep-fakes.[55]

Another example is the right to explanation and human intervention/oversight[56]. Article 22(3) of the GDPR mandates human intervention and provides a right to meaningful information for decisions based solely on automated processing, which also includes profiling under Article 15(1)(h). In contrast, the AI Act requires human oversight under Article 14, along with a right to explanation of individual decision-making in Article 86 specifically for high-risk AI systems. These regulations demonstrate significant differences in both content, prerequisites, and legal consequences.

Hence, under the provisions of the AI Act, it appears that, in principle, an AI system categorized as posing low or minimal risk may be permitted to make decisions that are entirely automated, as defined by the GDPR.[57] Based on this, the ECJ has ruled with its

50 Ibid.

51 Ibid.

52 M. WINAU, 'Areas of Tension in the Application of AI and Data Protection Law', 124.

53 The GDPR establishes the principle of transparency to facilitate the exercise of data subjects' rights under Article 15-22, including the right to erasure, to rectification and to data portability, whereas transparency obligations under the AI Act are imposed only for high-risk AI systems under Article 13 and for other certain AI systems under Article 50. Furthermore, Article 13 of the AI Act addresses the interests of the deployer of an AI system rather than those of the final user and/or data subject.

54 CEDPO AI Working Group, *AI and Personal Data A Guide for DPOs "Frequently Asked Questions"*, p. 11.

55 CEDPO AI Working Group, *AI and Personal Data A Guide for DPOs "Frequently Asked Questions"*, p. 11.

56 See also for the limitations of this provision S. WACHTER, B. MITTELSTADT, and L. FLORIDI, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2016).

57 'Key Issue 6: Interplay with GDPR - EU AI Act'.

landmark SCHUFA Holding judgment[58] "on credit scoring" and provided some guidance on the scope of the prohibition of automated decision-making under Article 22. According to this, the Court ruled that "*the assignment of automatically calculated credit scores is not in line with the GDPR, and the attribution of creditworthiness can already constitute a decision under Article 22 GDPR.*"[59]

2.2.2. The Three Main Overlapping Points

The two regimes have some overlaps, interacting most obviously with respect to ***(1) bias and discrimination,*** *(2)* ***risk assessments*** *and* ***(3) special categories of personal data***"[60]

(1) Bias and Discrimination:

Article 5(1)(a) of the GDPR establishes a broad requirement that personal data be processed fairly.[61] In the realm of personal data processing, fairness is an intentionally flexible principle, meant to encompass the many ways in which a party might process personal data unfairly.[62] However, in the realm of AI and machine learning, fairness is often more narrowly defined.[63] For instance, when evaluating the fairness of an AI or machine learning application as a Data Protection Officer (DPO), your primary concern would be the possibility of bias or discrimination inherent in the algorithms utilised.[64] This could involve racial bias, stemming from biased training data against a specific race, or gender bias, where one gender is disproportionately impacted. Such instances constitute the unfair handling of personal data.[65]

The right to data protection may be subordinate to the right to non-discrimination when the processing of special category data is strictly necessary to prevent bias.[66] It is indisputable that data protection should be upheld to the greatest extent possible. However, aside from the processing of

58 Case C-634/21 OQ v. Land Hesse, joined party: SCHUFA Holding AG (Scoring) Judgment of the Court (First Chamber) issued 7 December 2023, ECLI:EU:C:2023:957

59 'CJEU landmark rulings on "credit ranking" and review of DPAs'.

60 'Key Issue 6: Interplay with GDPR - EU AI Act'.

61 Article 5(1)(a), *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*.

62 CEDPO AI Working Group, *AI and Personal Data A Guide for DPOs "Frequently Asked Questions"*, p. 11.

63 CEDPO AI Working Group, *AI and Personal Data A Guide for DPOs "Frequently Asked Questions"*, pp. 11–12.

64 Ibid.

65 Ibid.

66 M. WINAU, 'Areas of Tension in the Application of AI and Data Protection Law', 133.

special categories of personal data to prevent biases, the extent to which the right to data protection should be limited to ensure the most accurate output of the system remains unclear.[67]

Under the EU AI Act, the concept of accuracy refers to AI systems ensuring their outputs are free from bias and do not result in discriminatory outcomes.[68] Therefore, in high-risk AI applications, accuracy includes not only the system's ability to generate correct results but also its compliance with non-discrimination and equality principles, ensuring that all individuals are treated regardless of their personal characteristics.

Article 10 of the AI Act addresses the issue of bias in training data, but it is specifically aimed at developers of predictive AI systems in high-risk areas, not developers of generative AI.[69] If generative AI is developed for a high-risk area, developers are required to comply with these rules. However, the effectiveness of Article 10 in addressing generative discrimination will largely depend on how 'bias' is interpreted within the AI Act.[70] Currently, it is uncertain whether the harms discussed in this chapter would be classified under the AI Act's concept of 'bias.'[71] Regulators might interpret 'bias' in a technical sense (regarding the diversity of training data) rather than in a social and ethical sense (relating to demeaning and abusive content).[72]

The Right to Explanation might not be useful for our purposes either. This provision focuses on explaining decisions made by predictive AI in high-risk areas like employment, criminal justice, or education.[73] Since the AI Act does not categorize generative AI as a high-risk application, the right to explanation generally does not apply to generative AI.[74] The AI Act does not include generally applicable provisions for developers of generative AI systems to mitigate bias.[75] The provisions concerning generative AI focus on transparency,

67 Ibid.

68 M. WINAU, 'Areas of Tension in the Application of AI and Data Protection Law', 130.

69 European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)) (Ordinary legislative procedure: first reading).References to articles and recitals introduced or changed by the Parliament will be labelled in this paper as "the EU AI Act".

70 P. HACKER, F. ZUIDERVEEN BORGESIUS, B. MITTELSTADT, and S. WACHTER, 'Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It' (2024) p. 39.

71 HACKER, ZUIDERVEEN BORGESIUS, MITTELSTADT, and WACHTER, 'Generative Discrimination', p. 39.

72 Ibid.

73 Article 86 European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(Ordinary legislative procedure: first reading).References to articles and recitals introduced or changed by the Parliament will be labelled in this paper as "the EU AI Act".

74 N. HELBERGER and N. DIAKOPOULOS, 'ChatGPT and the AI Act' (2023) 12 Internet Policy Review.

75 HACKER, ZUIDERVEEN BORGESIUS, MITTELSTADT, and WACHTER, 'Generative Discrimination', p. 40.

copyright protection, watermarking[76], and reporting of energy consumption.[77] Providers of generative AI models that pose systemic risks are required to conduct model evaluations, perform risk assessments (including red teaming), fulfill reporting obligations for serious incidents, and ensure cybersecurity measures are in place.[78] None of these obligations specifically target harms related to discrimination. It is considered that although risk assessments and mitigation efforts should encompass bias and non-discrimination, it remains to be seen whether the harmonized standards (Articles 40 and 41) and the codes of practice (Article 56) will effectively tackle discrimination issues in practice.[79] These standards will determine which types of bias are covered and the remedies that must be implemented.[80] To summarize, the essential provisions of the AI Act are aimed at non-generative AI, and regrettably, there are no explicit requirements imposing developers of generative AI models to design systems that mitigate bias.[81]

(2) **Risk Assessment:**

A Data Protection Impact Assessment (DPIA) concentrates on how risks related to personal data are alleviated, whereas a Fundamental Rights Impact Assessment (FRIA) has a broader scope, assessing the degree of impact on a variety of fundamental rights such as freedom of expression, access to justice, and the right to good administration.[82] This distinction is highlighted by the controllers' obligation under Article 25(1) of the GDPR to implement

76 Article 53(1)(a) and 55, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(Ordinary legislative procedure: first reading).References to articles and recitals introduced or changed by the Parliament will be labelled in this paper as "the EU AI Act".

77 AnnexXI, SectionII(2)(e) European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(Ordinary legislative procedure: first reading).References to articles and recitals introduced or changed by the Parliament will be labelled in this paper as "the EU AI Act".

78 Article55(a)and(d), European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(Ordinary legislative procedure: first reading).References to articles and recitals introduced or changed by the Parliament will be labelled in this paper as "the EU AI Act".

79 Hacker, Zuiderveen Borgesius, Mittelstadt, and Wachter, 'Generative Discrimination', p. 40.

80 Hacker, Zuiderveen Borgesius, Mittelstadt, and Wachter, 'Generative Discrimination', p. 40.

81 Hacker, Zuiderveen Borgesius, Mittelstadt, and Wachter, 'Generative Discrimination', p. 40.

82 A. Mantelero, 'The Fundamental Rights Impact Assessment (FRIA) in the AI Act: roots, legal obligations and key elements for a model template' (2024) p. 32; T. E. Loeber Lutz Riede, Christoph Werkmeister, Philipp Roos, Lena Isabell, 'EU AI Act unpacked #6: Fundamental rights impact assessment' (June 2024).

appropriate measures aimed at mitigating risks, exampling the GDPR's proactive stance in data processing[83]

The AI Act ensures the protection of fundamental rights through a technical way by banning AI systems with unacceptable risks, whereas the GDPR has a case-by-case basis approach in its evaluation of personal data processing, and the protection of personal data.[84]

More precisely, the GDPR aims to balance collective and individual interests in personal data processing, while protecting data subjects' rights by establishing provisions to be weighed on a case-by-case basis.[85] This balance is exemplified in the GDPR's risk mitigation tools.[86] A case-by-case basis approach is particularly relevant in how organisations conduct DPIA for any data processing high risk to the rights and freedoms of individuals.

However, the criteria and thresholds used to define high-risk AI systems under the AI Act may not align with those governing high-risk data processing under the GDPR, potentially resulting in divergent risk assessments and overlapping compliance obligations between the two regulations.[87]

It is important to highlight that a DPIA conducted pursuant to Article 35 of the GDPR complements a FRIA conducted under Article 27 of the EU AI Act. Consequently, a deployer may be considered as having fulfilled some obligations of the EU AI Act if they have already been addressed in a DPIA conducted under the GDPR. However, any specific requirements for a FRIA that have not been addressed in the DPIA must still be complied with by the deployers.

(3) Special Categories of Personal Data:

Article 9 of the GDPR concerns to the processing of special categories of personal data,[88] which can only be justified under specific circumstances such as explicit consent or legitimate purposes. "*However, there has been legal debate around the word "revealing", which suggests*

83 Winau, 'Areas of Tension in the Application of AI and Data Protection Law', 127.

84 Winau, 'Areas of Tension in the Application of AI and Data Protection Law', 127.

85 Winau, 'Areas of Tension in the Application of AI and Data Protection Law', 126.

86 Winau, 'Areas of Tension in the Application of AI and Data Protection Law', 127.

87 A. Schwanke, 'The EU AI Act: Summary and Key Issues' (September 2024).

88 Article 9(1):Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*.

that special category data is broader than just information about a person's racial or ethnic origin."[89] The European Court of Justice (CJEU) decided in the Case C-184/20[90] that "if an organisation can infer or deduce special category data, the information supporting that inference should also be treated as special category data In a machine learning context."[91] This ruling implies that proxy variables may be qualified as special category data under the GDPR.[92]

When it comes to generative AI, since it typically generates outputs unrelated to individuals' personal information, it is unlikely that this provision would apply to generative AI. For instance, if content produced by generative AI is racist or homophobic but not specifically tied to an individual, it would not fall under the scope of the GDPR. However, if the content is related to an individual, Article 9 of the GDPR may be breached, as providers often cannot invoke the exemptions outlined in Article 9(2) of the GDPR.

The EU AI Act, specifying the relationship of the data minimization principle and data governance obligations under its Article 10(5), stipulates that the providers of AI systems may exceptionally process special categories of personal data, but only to the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection, and correction related to these high-risk AI systems, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons, such as pseudonymization or encryption. According to Article 6(1)(f) of the GDPR, legitimate interest might serve as a lawful ground for processing to avoid biased data and discrimination. However, this legitimate interest alone is not sufficient, as it must also meet the exemption requirements under Article 9(2) of the GDPR, which is not always feasible[93] Article 10(5) of the EU AI Act might be invoked as an exemption under Article 9(2)(g) of the GDPR; however, ensuring safeguards for the data subject's fundamental rights could be challenging, especially when processing large amounts of special categories of data.

3. CONCLUSIONS

Overall, the EU AI Act and the GDPR follow similar regulatory objectives, aiming to protect individuals' rights and ensure ethical practices in technology and data use however, there are some tensions between them that need careful navigation. In cases of overlap, statutory provisions must be interpreted and applied with special consideration. However, such an interpretation may result in different understandings and bring about legal uncertainty, which could compromise the applicability of these two regulations. Therefore, a harmonized guideline is necessary for coherent and effective application.

89 'Key Issue 6: Interplay with GDPR - EU AI Act'.

90 ECJ, Case C-184/20, ECLI:EU:C:2022:601

91 'Key Issue 6: Interplay with GDPR - EU AI Act'.

92 'Key Issue 6: Interplay with GDPR - EU AI Act'.

93 'International: The interplay between the AI Act and the GDPR - AI series part 1'.

This legal uncertainty can pose challenges for organizations trying to comply with both sets of regulations simultaneously. For example, an AI system that processes personal data must comply with GDPR's strict data protection requirements, while also meeting the AI Act's mandates for risk management and transparency. Without clear guidelines, companies may struggle to balance these requirements, leading to inconsistent applications of the law and potential legal risks.

Therefore, a harmonized guideline is essential for a coherent and effective application of both the GDPR and the AI Act. Such a guideline should offer clear directions on how to align the obligations under both frameworks, ensuring that organizations can achieve compliance without unnecessary complexity. It should also address specific scenarios where conflicts might arise, providing practical solutions to reconcile the two regulations.

A unified approach would not only facilitate compliance but also enhance the protection of individuals' rights by ensuring that AI systems are developed and used in a manner consistent with the principles of data protection and ethical AI. This would promote greater legal certainty and foster trust in both regulatory frameworks, ultimately benefiting all stakeholders involved.

4. REFERENCES

Avocats, H., 'Actualités juridiques et digitales' https://info.haas-avocats.com/droit-digital, accessed 15 June 2022

CEDPO AI Working Group, *AI and Personal Data A Guide for DPOs "Frequently Asked Questions"* (2023)

'CJEU landmark rulings on "credit ranking" and review of DPAs', https://noyb.eu/en/cjeu-landmark-rulings-credit-ranking-and-review-dpas, accessed 8 July 2024

CLARK, J, DEMICRAN, M, KETTAS, K., 'Europe: The EU AI Act's relationship with data protection law: key takeaways' April 2024, https://privacymatters.dlapiper.com/2024/04/europe-the-eu-ai-acts-relationship-with-data-protection-law-key-takeaways/, accessed 1 July 2024

EBERS, M., 'Standardizing AI - The Case of the European Commission's Proposal for an Artificial Intelligence Act' (2021)

EBERS, M., 'Truly Risk-Based Regulation of Artificial Intelligence - How to Implement the EU's AI Act' (2024)

EBERS, M., V. R. S. HOCH, F. ROSENKRANZ, H. RUSCHEMEIER, and B. STEINRÖTTER, 'The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)' (2021) 4 *J* 589–603

'EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) | European Data Protection Board', https://www.edpb.europa.

eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-52021-proposal_en, accessed 4 July 2024

'EDPS Opinion 44/2023 on the Proposal for Artificial Intelligence Act in the light of legislative developments | European Data Protection Supervisor', September 2024, https://www.edps.europa.eu/data-protection/our-work/publications/opinions/2023-10-23-edps-opinion-442023-artificial-intelligence-act-light-legislative-developments, accessed 6 July 2024

'Ethics guidelines for trustworthy AI | Shaping Europe's digital future', April 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, accessed 2 April 2024

European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(Ordinary legislative procedure: first reading).References to articles and recitals introduced or changed by the Parliament will be labelled in this paper as "the EU AI Act",

HACKER, P., F. ZUIDERVEEN BORGESIUS, B. MITTELSTADT, and S. WACHTER, 'Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It' (2024)

HAJDUK, P., `AI ACT and GDPR: On the path towards overlap of the enforcement structures' October 2023, https://blog.ai-laws.org/ai-act-and-gdpr-on-the-path-towards-overlap-of-the-enforcement-structures/, accessed 6 July 2024

HELBERGER, N. and N. DIAKOPOULOS, 'ChatGPT and the AI Act' (2023) 12 *Internet Policy Review*

'International: The interplay between the AI Act and the GDPR - AI series part 1', November 2023, https://www.dataguidance.com/opinion/international-interplay-between-ai-act-and-gdpr-ai, accessed 30 June 2024

JACOBS, M. and J. SIMON, 'Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed By the European Commission' (2022) 1 *Digital Society* 6

'Key Issue 6: Interplay with GDPR - EU AI Act', https://www.euaiact.com/key-issue/6, accessed 18 June 2024

KUNER, C., F. H. CATE, O. LYNSKEY, C. MILLARD, N. NI LOIDEAIN, and D. J. B. SVANTESSON, 'Expanding the artificial intelligence-data protection debate' (2018) 8 *International Data Privacy Law* 289–92

LOEBER, T. E., RIEDE, L., WERKMEISTER, C., ROOS, P., ISABELL, L. 'EU AI Act unpacked #6: Fundamental rights impact assessment' June 2024, https://technologyquotient.freshfields.com//post/102j941/eu-ai-act-unpacked-6-fundamental-rights-impact-assessment, accessed 7 July 2024

MANTELERO, A., 'The Fundamental Rights Impact Assessment (FRIA) in the AI Act: roots, legal obligations and key elements for a model template' (2024)

Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle et modifiant les règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 et (UE) 2019/2144 et les directives 2014/90/UE, (UE) 2016/797 et (UE) 2020/1828 (règlement sur l'intelligence artificielle) (Texte présentant de l'intérêt pour l'EEE), (2024)

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), (2016)

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), (2024)

SCHWANKE, A., 'The EU AI Act: Summary and Key Issues' September 2024, https://medium.com/@axel.schwanke/eu-ai-act-summary-and-key-issues-531fbed12c97, accessed 7 October 2024

SOLIMANO, D. / P., 'Spain Just Created the First European AI Supervision Agency' August 2023, https://decrypt.co/153482/spain-just-created-the-first-european-ai-supervision-agency, accessed 6 July 2024

SPIECKER, I. and G. DÖHMANN,, 'AI and Data Protection' in L. A. DiMatteo, C. Poncibò, M. Cannarsa (eds.), The Cambridge Handbook of Artificial Intelligence, (Cambridge University Press, 2022), pp. 132–45

VEALE, M. and F. Z. BORGESIUS, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach' (2021) 22 *Computer Law Review International* 97–112

WACHTER, S., MITTELSTADT, B., and FLORIDI, L. 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2016)

WALLACE, N. and CASTRO, D., 'The Impact of the EU's New Data Protection Regulation on AI' March 2018, https://datainnovation.org/2018/03/the-impact-of-the-eus-new-data-protection-regulation-on-ai/, accessed 17 June 2024

WINAU, M., 'Areas of Tension in the Application of AI and Data Protection Law' (2023) 9 *European Data Protection Law Review* 123–35

Case C-184/20 Vyriausioji Tarnybines Etikos Komisija, issued 1August 2022, ECLI:EU:C:2022:601

Case C-634/21 OQ v. Land Hesse, joined party: SCHUFA Holding AG(Scoring) Judgment of the Court (First Chamber) issued 7 December 2023, ECLI:EU:C:2023:957

Capítulo 4

ENFORCING AI REGULATION IN FRANCE: A LEGAL FRAMEWORK BEYOND THE AI ACT

Sébastien Fassiaux
PhD Candidate
Universitat Pompeu Fabra

ABSTRACT: Artificial intelligence (AI) is sometimes perceived as a new, unregulated technology, despite its continued use in key markets throughout the development of the internet. This contribution argues that, beyond the recent hype around generative AI and the adoption of the EU's Artificial Intelligence Act, AI systems have long been subject to important legal limitations in Europe. First, it highlights the extent to which data protection rules have been regulating the use of AI systems, even before the application of the General Data Protection Regulation (GDPR). It also explores how competition and copyright laws complement data protection rules to further regulate AI systems. The article argues that national data protection authorities have been playing a major role in regulating algorithmic practices, long before the recent public interest for AI. Second, it discusses the extent to which the AI Act brings novelties to further regulate AI systems developed and deployed by private and public entities. It also examines the interactions between the AI Act and the GDPR. Third, focusing on the case of France, it explores how the country's data protection and competition authorities have contributed to algorithmic regulation over time, handling important cases related to AI systems in the fields of digital advertising, facial recognition, and generative AI.

KEYWORDS: Regulation; artificial intelligence; data protection; GDPR; Artificial Intelligence Act.

1. INTRODUCTION

Artificial intelligence (AI) is sometimes perceived as a new, unregulated technology. This misconception is mainly due to generative AI making the headlines since the introduction of ChatGPT in late 2022. The new chatbot created by OpenAI has been one of the fastest growing consumer products of all times.[1] Its release has clearly marked an inflexion point in consumers' awareness of AI's potential risks and opportunities. As a result, the EU legislator introduced provisions related to general purpose AI models in the newly enacted Artificial Intelligence Act.[2] This contributed to the new regulation giving the impression that it was the first legislative instrument applicable to AI systems.

However, AI systems are not being created and implemented in a legal void. In fact, before the hype around its generative form, AI has been operating in the background of the internet for long – powering search engines, personalised advertising, content recommendations, maps directions, etc. –, without consumers specifically realising that AI was at play. Due to their dependency on (personal) data, AI systems have been subject to data protection laws for some time now, even before the entry into force of the EU's General Data Protection Regulation (GDPR).[3] Today, arguably the most important limits to the use of AI systems are set by data protection and competition authorities.

If generative AI has now brought AI to the forefront of public debate, the EU had already been working on enacting the AI Act before ChatGPT's release. In this context, exploring the interactions between the AI Act and the GDPR is relevant because the former will be fully applicable only in 2027,[4] and its effects will take even longer to materialize. More generally, a better understanding of the overlap between AI and privacy policy issues appears essential for the future of AI development.[5]

---

1 The chatbot has gained 1 million users after a week and 100 million within two months (MILMO, Dan, "ChatGPT reaches 100 million users two months after launch", The Guardian, 2 February 2023, https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app).

2 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), OJ L, 12.7.2024, p. 1-144.

3 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1-88.

4 The AI Act entered into force on 1 August 2024. Most rules will be applicable as of 2 August 2026, except for (i) the prohibitions, definitions and provisions related to AI literacy (applying on 2 February 2025); (ii) the rules on governance and the obligations for general purpose AI (applying on 2 August 2025); and (iii) the obligations for high-risk AI systems that classify as high-risk because they are embedded in regulated products (applying on 2 August 2027). See Article 113 AI Act.

5 OECD, "AI, data governance and privacy: Synergies and areas of international co-operation", OECD Artificial Intelligence Papers, No. 22, Paris: OECD Publishing, 2024, https://doi.org/10.1787/2476b1a4-en.

Therefore, this contribution first highlights the extent to which data protection regulation already regulates the use of AI systems. It also explores how competition and copyright laws further complement privacy rules in AI regulation. Second, it discusses the extent to which the AI Act brings novelties to further regulate AI systems developed and deployed by private and public entities and examines the interactions between the AI Act and the GDPR. Third, the analysis focuses on the case of France, where the data protection and competition authorities have been handling important cases related to AI systems mainly in the fields of digital advertising, facial recognition, and generative AI.

2. BACKGROUND ON AI REGULATION

The fact that new legislation aiming to regulate the use of AI systems is being adopted at the EU level does not mean that the law disregarded the issue previously. Tech companies have been using AI systems in consumer markets since the early days of the internet. AI systems powering search engines, personalised advertising, and recommender systems made the fortunes of Google and Facebook over the past few decades. These companies are now among the most valuable on the planet. Yet, their use of AI systems for making huge quantities of information readily available and increasingly personalised has not been left unchecked. If the focus has mostly been on the regulation of AI through data protection law (Section 2.1), competition and intellectual property law is becoming increasingly relevant (Section 2.2).

2.1. AI Regulation Under Data Protection Law

Data protection law has initially been used to regulate AI systems. This is not surprising given their dependency on data for their training and operations. Important cases actually predate the GDPR and relate to fundamental questions over the balance between the right to privacy, the right to commercial exploitation of personal data, and the freedom of expression and information.

The issue of de-referencing is emblematic of this balance. In the landmark *Google Spain* case of 2014, the Court of Justice of the EU (CJEU) established the right to be forgotten in the context of search engine results.[6] The CJEU did so by interpreting the 1995 Data Protection Directive,[7] in light of the fundamental rights to privacy and to the protection of personal data guaranteed by the Charter of Fundamental Rights of the EU. This case raised the issue of an individual's right to be de-referenced from a search engine's results. Ultimate-

6 Judgment of 13 May 2014, Google Spain, C-131/12, ECLI:EU:C:2014:317.

7 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995, p. 31-50.

ly, the CJEU prioritized the fundamental right to privacy over the economic interests of the operator of an AI-powered search engine and the general public's interest in having access to personal information, subject to safeguards when the relevant person plays a role in public life. The judgment also highlighted the CJEU's role as a prominent guardian of individuals' privacy in Europe and the right it established was codified in the GDPR a few years later.

This case shows that important legal questions around AI have been in the open for some time now and did not appear only in the advent of the GDPR or – a fortiori – the AI Act.[8] Since *Google Spain*, the CJEU has continued to adjudicate important questions related to AI systems based on data protection law.[9]

As these pre-GDPR cases show, data protection requirements related to AI systems have thus preceded the recent adoption of the AI Act, the Digital Services Act (DSA),[10] or the Digital Markets Act (DMA),[11] which have established new rules and standards in this space. At any rate, the tension between the fundamental right to privacy and the commercial interests of large companies using AI systems is very much alive today, six years after the GDPR has started to apply.

Although early GDPR enforcement has been slow, it is now increasingly biting. Figure 1 below shows the highest fines issued under the GDPR as of August 2024. All of these fines were imposed on companies heavily relying on huge quantities of personal data for their AI systems driving personalised advertising, as well as content and product recommendations, among other things. Overall, the effectiveness of those fines and accompanying injunctions to deter tech firms from engaging in GDPR infringements can be discussed. But their general

8 Indeed, the plaintiff in the Google Spain case had complained to the Spanish data protection authority in 2010, almost 15 years ago and six years before the adoption of the GDPR.

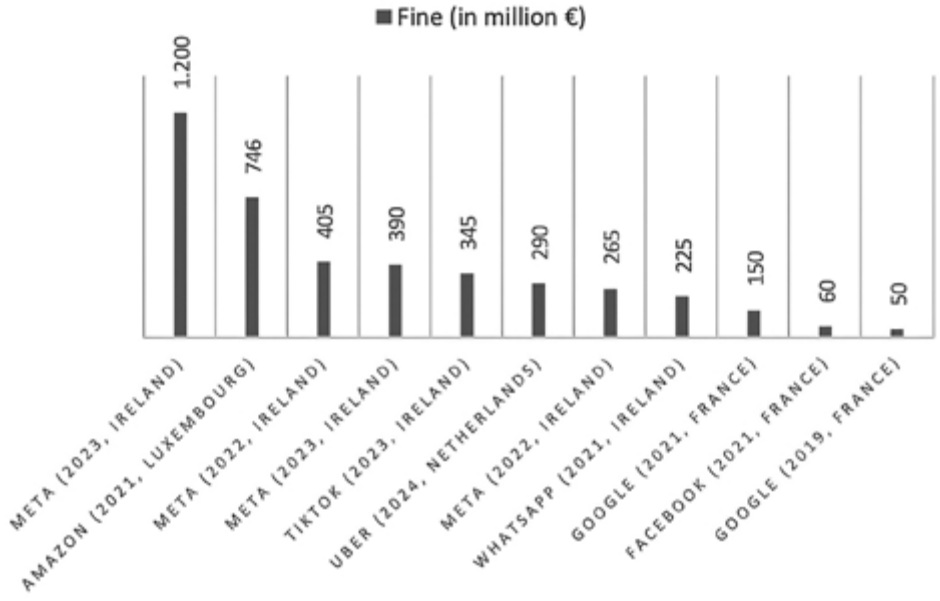
9 For example, still with regards to Google's search engine, the CJEU has also established in 2019 a "right to be accurately remembered" (EUROPEAN COMMISSION: LEGAL SERVICE, 70 years of EU law: a Union for its citizens, Luxembourg: Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2880/02622, p. 119), based on the GDPR read in light of the Charter of Fundamental Rights of the EU (judgment of 24 September 2019, GC and Others v CNIL, C-136/17, ECLI:EU:C:2019:773). Going further than pure de-referencing, this right can compel search engines to modify the ranking of their searches, directly affecting their search algorithms. In 2019 still, the CJEU delimited the territorial scope of the right to be de-referenced from a search engine's results (judgment of 24 September 2019, Google v CNIL, C-507/17, ECLI:EU:C:2019:772). The case originated in France, where the CNIL had imposed a fine of €100,000 on Google for refusing to de-reference results involving personal data to all worldwide versions of its search engine. The CJEU found that when a search engine like Google grants a request for de-referencing, it is in principle required to carry out the de-referencing with regards only to versions available in the EU, not elsewhere. The French Conseil d'Etat, which had referred the question to the CJEU, finally annulled the CNIL's fine to Google in 2020 (Conseil d'Etat, décision n°399922 du 27 mars 2020, ECLI:FR:CECHR:2020:399922.20200327).

10 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277, 27.10.2022, p. 1-102.

11 Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ L 265, 12.10.2022, p. 1-66.

influence on industry practices with regards to the handling of personal data elevates data protection law at the forefront of AI regulation.

Figure 1: Highest fines imposed under the GDPR as of August 2024



Sources: AP, CNIL, CNPD, DPC

Despite all of the affected companies having their EU establishments in Ireland, Luxembourg, and the Netherlands – thus giving data protection authorities there a preponderant influence in enforcing the GDPR due to its one-stop-shop mechanism –, the graph shows the active role of the CNIL, which imposed three of the highest fines under the GDPR.[12] The CNIL's enforcement action will further be discussed in Section 4 below.

12 Even when not directly fining undertakings, the CNIL has played an important role in some of the cases mentioned here by cooperating with the lead data protection authorities. For instance, it closely cooperated with the Dutch data protection authority regarding Uber's illegal data transfers to the United States, leading to a fine of €290 million (AUTORITEIT PERSOONSGEGEVENS, "Dutch DPA imposes a fine of 290 million euro on Uber because of transfers of drivers' data to the US", 26 August 2024, https://www.autoriteitpersoonsgegevens.nl/en/current/dutch-dpa-imposes-a-fine-of-290-million-euro-on-uber-because-of-transfers-of-drivers-data-to-the-us).

2.2. AI Regulation Under Competition and Intellectual Property Law

Although the focus of this contribution is on exploring the interplay between the GDPR and the AI Act, it is also relevant to note that AI regulation is increasingly taking place in other fields, including competition and intellectual property law. The scope and objectives of all three areas differ substantially, though.

The main objective of data protection law is to protect individuals' privacy, and its scope is rather broad, as exemplified by the AI-related cases in France discussed in Section 4. Being technology-neutral and applying horizontally to both private and public entities, the GDPR applies to a wide range of practices involving AI systems. Its comprehensive rules and principles leave sufficient room for interpretation, which is both its strength and weakness.

Instead, the main objective of competition law in this context is to ensure that AI-intensive markets remain fair and contestable, guaranteeing a level-playing field for businesses, ultimately benefitting consumers. Here, trustbusters have to strike a difficult balance between innovation and competition on the merits. Traditional competition enforcement based on Articles 101 and 102 of the Treaty on the Functioning of the EU has led to important fines, also in AI-related markets.[13] Yet, traditional, ex post competition rules have been difficult to enforce in digital markets. Proceedings take years to result in sanctions and, in the meantime, harm to competitors and consumers is irreversible.

In order to reverse this trend, the EU legislator adopted the DMA in 2022. Designed as an ex-ante regulation mainly concerned with ensuring competitive markets, the DMA complements traditional competition rules by applying more specific obligations in a number of defined digital markets controlled by gatekeepers designated as such by the Commission.

While DMA enforcement is still in its infancy, some experts already argue that, contrary to the AI Act, the DMA contains "the most far-reaching, most overlooked but potentially also most effective regulatory constraints for AI."[14] What is sure is that the DMA will

13 For instance, the largest-ever competition fine of €4.34 billion was imposed by the European Commission on Google in 2018 (case AT.40099 – Google Android) – it was later reduced to €4.125 billion by the CJEU (judgment of 14 September 2022, Google Android, T-604/18, ECLI:EU:T:2022:541). The Commission had found that Google had abused its dominance in the market for mobile operating systems with Android to safeguard its dominance on the market for general search services, thus protecting its main source of revenue from its AI-driven search engine.

14 HACKER, Philipp; CORDES, Johann; ROCHON, Janina, "Regulating Gatekeeper Artificial Intelligence and Data: Transparency, Access and Fairness under the Digital Markets Act, the General Data Protection Regulation and Beyond", European Journal of Risk Regulation (2023) 15(1), p. 51. At a recent workshop on generative AI, EU Commissioner for Competition Margrethe Vestager also confirmed that the DMA applied to generative AI features embedded in services covered by the regulation: "We continue to apply our trusty merger and antitrust rules. Even though we sometimes have to remind AI market players that the competition rules also apply to them. And our Digital Markets Act applies too: the DMA can also regulate AI even though it is not listed as a core platform service itself. AI is covered where it is embedded in designated core platform services such as search engines, operating systems and social networking services. So we are applying our rulebook to concerns

profoundly affect AI-related markets and provide enforcers with additional tools to regulate them.[15]

At the same time, discussions around copyright are gaining importance in the context of the development of generative AI applications and their underlying foundation models. Indeed, such models require training on very large datasets, but developers have been accused of training them on copyrighted materials. In this regard, the 2019 Directive on Copyright in the Digital Single Market (CDSM)[16] allows reproductions and extractions for text and data mining under specific conditions.[17] What is at stake is the tension between the ability to innovate and the protection of authors. The AI Act follows this approach by requiring providers of general-purpose AI models to put in place policies to comply with EU copyright rules, including with the text and data mining exemption under the CDSM.[18]

This is particularly relevant for the news media, which have first had to adapt to the development of the internet, then to digital platforms, and now see a threat from generative AI applications. In fact, the French competition authority has imposed a fine of €250 million on Google in March 2024 for breaching commitments in a case where the tech company was found to have infringed the French law transposing the CDSM and aimed at ensuring fair negotiations between online platforms, press agencies, and publishers.[19] The regulator also found that Google's AI service Bard (now renamed Gemini) utilised content from press agencies and publishers without notification or opt-out options, hindering their ability to negotiate fair remuneration.[20] This case highlights how competition and copyright issues are equally relevant for AI regulation.

---

that we have already in the AI world. We are looking at the issues very closely, from all angles and with all our tools" (EUROPEAN COMMISSION, "Speech by EVP Margrethe Vestager at the European Commission workshop on 'Competition in Virtual Worlds and Generative AI'", 28 June 2024, https://ec.europa.eu/commission/presscorner/detail/en/speech_24_3550).

15 For instance, after the German competition authority had innovated in 2019 by considering personal data handling by dominant firms, finding Meta had inappropriately combined personal data across its AI-driven services (Bundeskartellamt, decision B6-22/16 of 6 February 2019), the DMA now also considers such conducts by gatekeepers (Article 5(2)(b) DMA).

16 Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, OJ L 130, 17.5.2019, p. 92-125.

17 Under Article 4(3) CDSM, rightsholders can reserve their rights to prevent such mining unless it is for scientific research. If rights are reserved, AI model providers must obtain authorisation from rightsholders for text and data mining. This possibility for rightsholders to opt-out from the exemption of copyright protection has been criticised for going precisely beyond copyright protection. Instead, by allocating property rights for AI's "building blocks" the EU legislator appears to have adopted a "property-right approach to the regulation of AI" (MARGONI, Thomas; KRETSCHMER, Martin, "A deeper look into the EU text and data mining exceptions: Harmonisation, data ownership, and the future of technology", GRUR International (2022) 71(8), p. 688).

18 Article 53(1)(c) AI Act.

19 Autorité de la concurrence, décision 24-D-03 du 15 mars 2024.

20 This fine represents more than what the CNIL has ever imposed in a single year for GDPR infringements, to date.

3. THE AI ACT AND ITS INTERPLAY WITH THE GDPR

The AI Act integrates with the strategy outlined in the European Commission's 2020 White Paper on Artificial Intelligence by operationalising its vision of an ecosystem of excellence and trust.[21] In line with the objective of fostering trustworthy AI, the Commission highlighted the need to protect the fundamental right to privacy in the context of AI regulation.

After the Commission made a proposal for an AI Act in 2021, it was finally approved by the Council in May 2024. The AI Act is the first comprehensive set of rules specifically applying to AI systems worldwide. It entered into force on 1 August 2024. These rules are the result of intense negotiation among the Commission, Parliament, and Council. With around 3,000 amendments, it has probably been the most debated legislative file of the 2019-2024 legislature.[22] While the Commission aims at promoting the AI Act as a global standard for AI systems regulation, a relevant question is whether it will effectively play that role, just like the GDPR for privacy, or if its Brussels effect[23] will be more limited.

The Act complements the data protection, competition, and copyright law regimes applicable to AI systems. It is mainly a "products" regulation and aims to ensure AI systems placed on the market are safe – just like any other products[24] – and respect fundamental rights. It

21 EUROPEAN COMMISSION, White Paper on Artificial Intelligence - A European approach to excellence and trust, 19.2.2020, COM(2020) 65 final. The White Paper stressed the necessity to establish a regulatory framework addressing high-risk AI systems by setting requirements for data governance, transparency, and cybersecurity, involving both pre-market evaluations and post-market monitoring. The EU's AI strategy is ultimately furthering the goals of its digital strategy, outlining policy measures to advance the digital transition with regards to skills, infrastructures, public services, and the economy (EUROPEAN COMMISSION, Shaping Europe's Digital Future, Luxembourg: Publications Office of the European Union, 2020, https://commission.europa.eu/system/files/2020-02/communication-shaping-europes-digital-future-feb2020_en_4.pdf; EUROPEAN COMMISSION, 2030 Digital Compass: the European way for the Digital Decade, 9.3.2021, COM(2021) 118 final).

22 Difficult negotiations took place particularly concerning foundational models. France partially opposed the regulation, concerned that it could hinder innovation and the development of European AI models aligned with local language and culture. Representing the views of liberal leaders, Jean-Noël Barrot explained as French Minister for Digital Transition in November 2023 that there was still hope for European models to develop in the coming years, but that regulation should not hinder innovation and rather focus on protecting consumers and citizens (CNIL, "Cahier air2023. IA et libre-arbitre: sommes-nous des moutons numériques?", 22 April 2024, https://www.cnil.fr/sites/cnil/files/2024-04/cahier_air2023.pdf, p. 8).

23 BRADFORD, Anu: The Brussels Effect: How the European Union Rules the World, New York: Oxford University Press, 2020.

24 Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC, OJ L 135, 23.5.2023, p. 1-51.

does so by categorizing AI systems by risk level and imposes obligations accordingly.[25] In practice, this means that most AI systems will not be heavily regulated.

The AI Act interacts with the GDPR in a number of ways. First of all, the AI Act specified that it does not alter existing EU laws on personal data processing or the duties of supervisory authorities overseeing compliance. Thus, it maintains the obligations of AI providers and deployers as data controllers or processors under EU or national data protection laws. Similarly, data subjects retain all rights under these laws, including those related to automated decision-making and profiling. However, the AI Act complements the GDPR with regards to (i) data quality; and (ii) the competence of national data protection authorities.[26]

First, with regards to data quality, the Act mandates that high-risk AI systems be trained on high-quality data to ensure AI systems are safe, effective, unbiased and non-discriminatory, adhering to data protection laws like the GDPR.[27] However, the AI Act complements the GDPR because, contrary to the latter, its requirements are not only limited to personal data.

Yet, the AI Act foresees an exception to the strict GDPR rules on special categories of personal data.[28] Indeed, the GDPR in principle prohibits the processing of such data and only allows it in limited circumstances.[29] But the AI Act adds a new exception by giving the possibility for providers of high-risk AI systems to exceptionally process these sensitive data to

25 The regulation requires CE marking for high-risk uses, achieved through the compliance with significant obligations for AI systems providers, deployers, importers, and distributors, controlled by rigorous audits (Article 6-49). General purpose AI models are generally subject to limited obligations, for instance related to documenting their sources of training, and respect for copyright (Article 51-54), except those involving systemic risks based on their increased capabilities which are also subject to significant obligations (Article 55). But low-risk AI systems face lighter requirements, for example in terms of transparency (Article 50). Yet, in order to protect fundamental rights and European values, some AI systems involving unacceptable risks are prohibited outright (Article 5).

26 Another one relates to regulatory sandboxes. The AI Act allows AI providers to use personal data lawfully collected for another purpose to develop, train, or test certain AI systems in those sandboxes, but only if they serve public interests (e.g., public safety, public health, environment, energy sustainability, transports, or public administration). In order to avoid inconsistencies with the GDPR (EDPB-EDPS, "Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)", 18 June 2021, https://www.edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf), for example the principle of purpose limitation, the AI Act outlines strict conditions under which this personal data can be used within regulatory sandboxes. For instance, it foresees safeguards for sensitive data and explicitly provides that this possibility should not constitute an exemption to the right not to be subject to a decision based solely on automated processing, including profiling (Article 22 GDPR).

27 Article 10 AI Act. Data used for AI training and testing must be relevant, representative, complete, and error-free, with clear transparency on data collection purposes. Privacy-preserving techniques should be used, and data should reflect specific usage contexts.

28 That is, personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for unique identification, data concerning health, and data regarding a natural person's sex life or sexual orientation (Article 9(1) GDPR).

29 Article 9(1) GDPR.

detect and correct bias in their datasets, if strictly necessary.[30] However, their use is subject to strict conditions. Despite being an exception to the prohibition of sensitive data processing, this possibility should be welcomed to guarantee that AI systems are trained and tested on unbiased datasets, maximising the respect of fundamental rights when deployed.

Second, the AI Act extends the competence of national data protection authorities. As anticipated above, the AI Act puts in place a system of post-market monitoring. Under this system, national market surveillance authorities are tasked with controlling compliance with the AI Act,[31] except for general purpose AI models for which the regulation grants exclusive competence to the European Commission,[32] whose newly established AI Office will enforce rules and monitor compliance. Of course, Member States already have market surveillance authorities to monitor compliance with product safety rules. But the Act specifies that, for high-risk AI systems used in law enforcement, border control management, and the administration of justice and democratic processes – including the use of biometrics in these three areas –, Member States need to designate their supervisory authorities under either the GDPR or the Law Enforcement Directive[33] for market surveillance and control.[34] In France, the CNIL will thus have extended competences under the AI Act. This might create additional bottlenecks for national data protection authorities which already suffer from a lack of resources.[35]

4. REGULATING AI THROUGH DATA PROTECTION ENFORCEMENT IN FRANCE

Given the decentralised nature of the enforcement system under the GDPR, each Member States has designated a supervisory authority to enforce it. France designated the *Commission nationale de l'informatique et des libertés* (CNIL), which was established in 1978 as an independent administrative body to act as the country's data protection authority.[36]

30 Article 10(5) AI Act.

31 See Chapter III, Section 4; and Article 70 AI Act.

32 See Chapter V AI Act.

33 Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ L 119, 4.5.2016, p. 89-131.

34 Article 74(8) AI Act.

35 EU AGENCY FOR FUNDAMENTAL RIGHTS, "Lack of resources undermine EU data protection enforcement", 11 June 2024, https://fra.europa.eu/en/news/2024/lack-resources-undermine-eu-data-protection-enforcement.

36 Under the GDPR, it can now impose fines of up to €20 million or, for companies, up to 4% of their total worldwide annual turnover, whichever is higher (Article 83 GDPR).

For some years now – and before the hype around generative AI –, the CNIL has reflected on AI development and regulation.[37] In this context, the CNIL has long addressed AI issues through case studies, such as the use of smart cameras in public spaces, voice assistants, and facial recognition.[38]

At a CNIL event on AI and free will held in November 2024, its president Marie-Laure Denis explained that one could not wait for the AI Act to regulate AI systems and that the GDPR was sufficiently flexible to ensure positive AI development. As developed below, the CNIL indeed has a proven track record of active GDPR enforcement in different AI fields (Section 4.1) and has already started working on its application to generative AI, along with its European peers (Section 4.2).

4.1. Early GDPR Enforcement

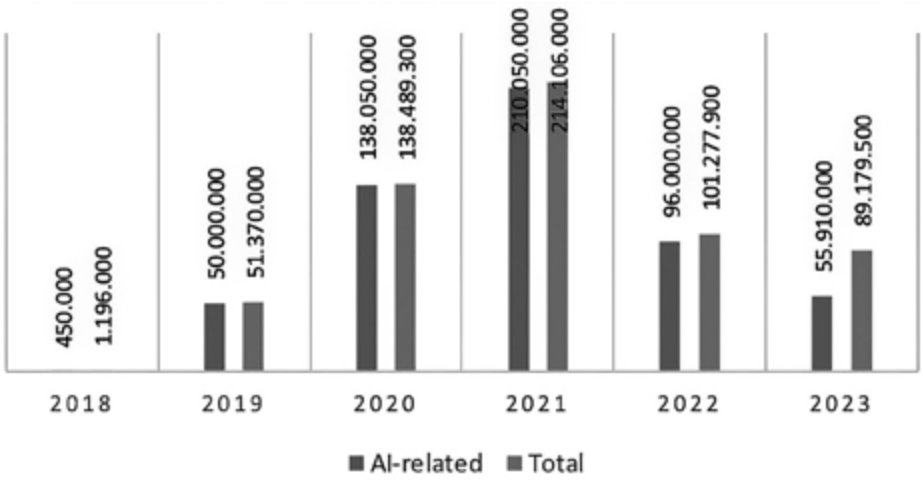
The CNIL has been one of the most active GDPR enforcers in Europe, also in the AI field. In fact, the GDPR did mark a breaking point in terms of sanctioning the mishandling of personal data (see Figure 2 below). Pre-GDPR, the French regulator was imposing modest fines.[39] Post-GDPR, that is after the regulation became applicable in May 2018, the regulator's fines reached record-high levels, with individual sanctions in the millions, even reaching €150 million for a fine imposed on Google in 2021.

37 Already in 2017, it published a report following a public debate on the ethical challenges of AI (CNIL, "Concertation citoyenne sur les enjeux éthiques liés à la place des algorithmes dans notre vie quotidienne: synthèse de la journée", 17 October 2017, https://www.cnil.fr/sites/cnil/files/atoms/files/cr_concertation_citoyenne_algorithmes.pdf). In its 2018 annual report – the year the GDPR began applying –, the CNIL dedicated a section to AI, mainly focusing on its ethical and societal implications (CNIL, "Rapport d'activité 2018", 15 April 2019, https://www.cnil.fr/sites/cnil/files/atoms/files/cnil-39e_rapport_annuel_2018.pdf). It emphasised the importance of educating the public about AI's potential and fostering a sustainable AI model at the national, European, and international levels. Moreover, it underscored the need for ethical considerations in AI, addressing issues like human autonomy, algorithmic discrimination, and the impact of hyper-personalisation on societal structures. Given the importance of personal data for training and operating AI systems, it also highlighted the regulatory role of the GDPR in creating a trustworthy framework for data use in AI, ensuring transparency and individual rights.

38 It has also reviewed government projects like the AI tools for tax fraud detection. Since 2022, the CNIL has published resources clarifying AI-related challenges and GDPR compliance, offering an AI self-assessment guide for the public and professionals.

39 In 2016, the CNIL imposed fines for a total amount of €160,000, and more than doubled them in 2017 (€371,000).

Figure 2: Share of CNIL fines to entities using AI systems, compared to total fines imposed (2018-2023)



Source: CNIL

Interestingly, a closer look at the nature of the data processing activities at hand shows that fines particularly affected entities heavily relying on personal data for their AI-powered services. In fact, the overwhelming majority of fines where imposed on companies for their AI-related activities (see Figure 2 above).

The CNIL has fined companies using AI systems in a number of domains, namely: data security, personalised advertising, facial recognition, and data brokerage. Table 1 below offers an overview of the main decisions issued by the CNIL related to activities involving the use of AI since the GDPR became applicable. Although the overview is non-exhaustive, it contains the most important decisions based on the fines involved. It is also representative of the focus given by the CNIL in its enforcement action, for instance with regards to personalised advertising.

Table 1: CNIL decisions sanctioning entities intensively using AI systems (2018-2024)

Date	Domain	Entity	Fine
04/04/2024	Data brokerage	Hubside.store	€525,000
31/01/2024	Data brokerage	Foriou	€310,000
29/12/2023	Data brokerage	Tagadamedia	€75,000
29/12/2023	Personalised advertising	Yahoo	€10 million
29/12/2023	Personalised advertising	NS Cards	€105,000
15/06/2023	Personalised advertising	Criteo	€40 million
08/06/2023	Personalised advertising	KG COM	€150,000
11/05/2023	Personalised advertising	Doctissimo	€380,000
17/04/2023	Facial recognition	Clearview AI	€5.2 million
29/12/2022	Personalised advertising	Voodoo	€3 million
29/12/2022	Personalised advertising	TikTok	€5 million
29/12/2022	Personalised advertising	Apple	€8 million
19/12/2022	Personalised advertising	Microsoft	€60 million
17/10/2022	Facial recognition	Clearview AI	€20 million
31/12/2021	Personalised advertising	Facebook	€60 million
31/12/2021	Personalised advertising	Google	€150 million
27/07/2021	Personalised advertising	Le Figaro	€50,000
07/12/2020	Personalised advertising	Amazon	€35 million
07/12/2020	Personalised advertising	Google	€100 million
18/11/2020	Personalised advertising	Carrefour	€3.05 million
21/01/2019	Personalised advertising	Google	€50 million
19/12/2018	Data security	Uber	€400,000
24/07/2018	Data security	Dailymotion	€50,000

Source: CNIL

As this overview shows, the CNIL has indeed given priority to enforcing the GDPR with regards to personalised advertising, which heavily relies on AI systems to operate. For instance, for a number of years, the CNIL has prioritised investigations related to cookies and dark patterns.[40]

In its administrative practice, the authority has relied on the latest research, also based on findings in computer science. In 2020, an empirical study showed that so-called "cookie

40 Dark patterns are deceptive designs of user interfaces leading users to accept things they would not otherwise have agreed to, such as the placing of cookies on their devices for personalised advertising purposes.

banners" appearing on websites and seeking consumer consent to place cookies on their devices were mainly not compliant with the GDPR: just 11.8% of websites using the five top consent management platforms met minimal GDPR requirements for valid consent.[41] The CNIL inter alia referred to this empirical study in its decision to fine Facebook €60 million for the use of dark patterns, namely not enabling users to reject cookies as easily as accepting them.[42]

The CNIL also fined Google €150 million for similar practices.[43] Applying the French law transposing the ePrivacy Directive[44] in light of the heightened consent requirements under the GDPR, the authority held that the method employed by Google Search and YouTube for users to manifest their choice over the placing of cookies was illegally biased in favour of consent. Again, the authority referred to several studies showing that organisations implementing a "refuse all" button on the first-level consent interface had seen a decrease in the consent rate to accept cookies. The CNIL also relied on the same studies to impose heavy fines on Microsoft[45] and TikTok.[46]

The CNIL focused not only on deceptive designs that manipulate users into sharing more personal information than they intended but also on other types of invalid consent or the lack thereof. For instance, the CNIL imposed a €50 million fine on Google in 2019 for failing to obtain valid consent (i.e., informed, unambiguous, and specific) from its users for purposes of personalised advertising,[47] later upheld by the *Conseil d'Etat*.[48] Since then, Google adapted its practices across the EU to comply with GDPR requirements.

41 NOUWENS, Midas; LICCARDI, Ilaria; VEALE, Michael; KARGER, David; KAGAL, Lalana, "Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence", Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (2020), pp. 1-13.

42 CNIL, délibération SAN-2021-024 du 31 décembre 2021. The CNIL found that discouraging users to decline cookies while encouraging them to accept being tracked on the first page undermined their freedom of consent, as many users would not accept cookies if offered a genuine choice.

43 CNIL, délibération SAN-2021-023 du 31 décembre 2021.

44 Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), OJ L 201, 31.7.2002, p. 37-47.

45 CNIL, délibération SAN-2022-023 du 19 décembre 2022.

46 CNIL, délibération SAN-2022-027 du 29 décembre 2022.

47 CNIL, délibération SAN-2019-001 du 21 janvier 2019. The regulator determined that user consent was not sufficiently informed because the information about the company's processing of personal data for ad personalisation was not centralised in a single document. As a result, users were unaware that their information would be processed for this purpose across Google's various services. Nor was consent unambiguous because the option to be shown personalised ads was pre-checked in difficult-to-find settings. Finally, nor was consent specific, because Google sought consent when users were creating an account, by ticking boxes to agree on its terms of service and privacy policy, whereas consent should be obtained distinctively for each purpose.

48 Conseil d'Etat, décision n°430810 du 19 juin 2020, ECLI:FR:CECHR:2020:430810.20200619.

More recently, the CNIL fined Yahoo €10 million for placing advertising cookies on users' devices without collecting their prior consent at all.[49] The CNIL found that requiring non-essential cookies for service use is not illegal if users can freely give or withdraw consent without negative consequences. However, the authority noted that, in Yahoo's case, withdrawal was difficult due to service interruptions, lack of alternatives, and misleading interface elements that complicated the process for users.

In total, just between 2020 and 2021, the CNIL adopted around 70 corrective measures (formal notices and sanctions) related to non-compliance with cookie regulation, especially related to consent.[50] In total, 80% of the affected entities rapidly complied with the authority's orders, with effects not only in France but at least across the EU.[51]

Overall, if advertising might not be the first thing popping to mind when thinking about AI, it has played a key role in how the digital economy has developed, with relevance in today's developments around generative AI. Advertising has been described as the "lifeblood of the internet"[52] because it has enabled business models offering free services, making possible the development of some of the largest online platforms, such as those operated by Google and Meta. Many of those platforms have now been designated as gatekeepers under the DMA because of their important market power and ability to raise barriers to entry for potential competition. In fact, most fines imposed by the CNIL in this space relate to the illegal placing of advertising cookies linked to both companies.

In turn, companies like Alphabet, Amazon, Microsoft, and Meta – which have relied on scores of personal data and engaged in personalised advertising – are now among the main developers of, and investors in generative AI models.[53] Therefore, the link between personalised advertising and further AI development is more important than usually believed, yet at the core of the activities of some of the most valuable companies on the planet.

Finally, the CNIL's early GDPR enforcement has also focused on another domain relevant for its interaction with the AI Act: that of facial recognition. In 2022, Clearview AI, which amassed over 20 billion images for its facial recognition service by scraping publicly accessible websites, was fined €20 million by the CNIL for multiple GDPR violations.[54] At any rate, the AI Act will now explicitly

49 CNIL, délibération SAN-2023-024 du 29 décembre 2023.

50 CNIL, "Refuser les cookies doit être aussi simple qu'accepter: bilan de la deuxième campagne de mises en demeure et actions à venir", 14 September 2021, https://www.cnil.fr/fr/refuser-les-cookies-doit-etre-aussi-simple-quaccepter-bilan-de-la-deuxieme-campagne-de-mises-en.

51 Ibid.

52 COFONE, Ignacio; ROBERTSON, Adriana, "Consumer Privacy in a Behavioral World", Hastings Law Journal (2018) 69(6), p. 1472.

53 AUTORITE DE LA CONCURRENCE, "Avis 24-A-05 du 28 juin 2024 relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative", https://www.autoritedelaconcurrence.fr/sites/default/files/integral_texts/2024-06/avisIA.pdf.

54 CNIL, délibération SAN-2022-019 du 17 octobre 2022. The investigation revealed Clearview's unlawful data processing without user consent, and failure to respect individual data rights of access and erasure. Clearview AI also failed to cooperate with the CNIL in this initial investigation, as a result of which it was ordered to cease

prohibit AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage,[55] such as those offered by Clearview AI.

4.2. Application of Data Protection Rules to Generative AI Systems

Since the release of advanced generative AI applications as of late 2022, the CNIL has started working on applying the GDPR to these new systems. So far, it has mainly focused on policy work, as no sanctions have yet been imposed on generative AI companies. For instance, the authority unveiled an AI action plan in 2023, extending its policy and regulatory work to generative AI and large language models.[56] Overall, its action plan reflects a comprehensive approach to balance innovation with privacy and ethical considerations.

However, France is not the only Member State where data protection authorities have initiated discussions and enforcement actions regarding generative AI applications under the GDPR. Important concerns relate to the very development of generative AI models. In May 2024, the Dutch data protection authority (AP) has issued guidelines related to web scraping, i.e. the automatic collection and storage of online information.[57] The authority determined that this practice is almost always illegal due to privacy risks and GDPR violations. This is particularly relevant for generative AI applications that rely on large datasets, as they often involve the collection of personal data without consent. The AP emphasised that publicly accessible information does not imply permission for scraping, and exceptions are rare, typically limited to non-commercial, personal projects or highly targeted corporate uses. These findings raise doubts as to the compatibility of web scraping involving personal data with the GDPR, but the issue is not settled yet.

In fact, the Italian data protection authority is still conducting investigations against OpenAI for alleged GDPR breaches,[58] following a temporary ban of its ChatGPT chatbot in

data collection in France and delete existing data, facing additional daily penalties for non-compliance, totalling €5,2 million in 2023 (CNIL, délibération SAN-2023-005 du 17 avril 2023).

55 Article 5(1)(e) AI Act.

56 CNIL, "Intelligence artificielle: le plan d'action de la CNIL", 16 May 2023, https://www.cnil.fr/fr/intelligence-artificielle-le-plan-daction-de-la-cnil. In its plan, the CNIL recognises the importance of personal data in this space, focusing on transparency, data security, bias prevention, and the ethical use of AI technologies. Building on years of prior work in the AI field, the CNIL's agenda expands to include generative AI, chatbots, and other derivative applications. The plan is structured around four key objectives: (i) understanding AI systems and their impacts; (ii) ensuring privacy-respecting AI development; (iii) fostering innovation within the AI ecosystem in France and Europe; and (iv) and auditing and regulating AI systems to protect individuals. The plan includes publishing guidelines (many of which have been subject to public consultation), engaging with AI developers, and conducting audits.

57 AUTORITEIT PERSOONSGEGEVENS, "AP: scraping bijna altijd illegaal", 1 May 2024, https://autoriteitpersoonsgegevens.nl/actueel/ap-scraping-bijna-altijd-illegaal.

58 Garante per la protezione dei dati personali, "ChatGPT: Garante privacy, notificato a OpenAI l'atto di contestazione per le violazioni alla normativa privacy", 29 January 2024, https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9978020.

March 2023.[59] The suspected violations include illicit personal data collection via web scraping without user consent or proper legal basis.

Privacy concerns have also arisen regarding the training of AI models based on user-generated content on social media. Meta's announcement in May 2024 that it was updating its privacy policy to allow for the training of its AI models with user data sparked controversy over GDPR compliance. In its updated policy, Meta claimed that it had legitimate interests to train its AI models on the content users generated on Facebook and Instagram, including personal data, thus justifying bypassing user consent for this type of data processing. However, following several GDPR complaints – including before the CNIL –, Meta postponed its AI features in Europe.[60] Putting pressure on regulators, Meta cited concerns over innovation and competitiveness on the continent.[61]

Further privacy issues relate to the use of generative AI systems when they have already been deployed. In its investigations into OpenAI, the Italian data protection authority is concerned that ChatGPT's lack of an age verification system exposes minors to inappropriate content.[62] In April 2024, the Austrian data protection authority received a complaint directed at OpenAI for the inherent inaccuracies and lack of transparency in data generated by ChatGPT.[63] Despite requests for data access and rectification, OpenAI contends that it cannot rectify generated data, thus potentially infringing upon GDPR provisions.[64]

These recent developments in regulatory scrutiny over generative AI highlight the need for a common European approach in AI regulation. The CNIL and other data protection authorities have indeed started cooperating in this space, both for the adoption of legislation[65] and in enforcement actions.[66]

59 Garante per la protezione dei dati personali, Provvedimento del 30 marzo 2023 [9870832].

60 META, "Building AI Technology for Europeans in a Transparent and Responsible Way", 10 June 2024, https://about.fb.com/news/2024/06/building-ai-technology-for-europeans-in-a-transparent-and-responsible-way/.

61 Ibid.

62 Other concerns relate to ChatGPT's tendency to produce false information ("hallucinating"), particularly regarding individuals, raising significant implications for GDPR's obligations on personal data accuracy and user rights in this regard.

63 NOYB, "ChatGPT provides false information about people, and OpenAI can't correct it", 29 April 2024, https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it.

64 Ibid.

65 In 2018, the CNIL already highlighted that it was actively involved in shaping international guidelines and ethical standards for AI development (CNIL, "Rapport d'activité 2018", 15 April 2019, https://www.cnil.fr/sites/cnil/files/atoms/files/cnil-39e_rapport_annuel_2018.pdf, p. 31). A few years later, it was involved in shaping the AI Act based on the European Commission's proposal of 2021. Indeed, the authority collaborated with its European peers within the European Data Protection Board (EDPB) to assess the proposal and make recommendations. This cooperation resulted in the publication of an opinion in which data protection authorities highlighted the important overlaps between AI and data protection regulation and the challenges in aligning the AI Act with the GDPR (EDPB-EDPS, "Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)", 18 June 2021, https://www.edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf).

66 Given the many privacy issues arising from the development and deployment of generative AI systems, various data protection authorities within the EDPB have established a taskforce dedicated to interpreting the GDPR

5. CONCLUSIONS

This contribution argues that, beyond the recent hype around generative AI and the adoption of the AI Act, AI systems have long been subject to important legal limitations. This has been the case in France and across Europe.

Given the centrality of (personal) data for the development and use of AI systems, data protection rules have been instrumental in regulating them. As discussed here, national data protection authorities such as France's CNIL have been playing a major role in regulating algorithmic practices. Under the AI Act, their competence will be expanded but more resources are probably needed for effective enforcement. In addition, competition and copyright laws have also, and will continue to substantially impact the development and use of AI systems.

Yet, regulatory gaps remain and are sometimes filled by courts. In France, the *Conseil constitutionnel* has even applied the Declaration of the Rights of Man and of the Citizen of 1789 to fill such a gap. Establishing a constitutional right to access administrative documents, the court specified how universities must reveal the criteria and potential algorithmic methods used to evaluate student applications.[67] Despite the enthusiasm around the AI Act, legislation is therefore not a panacea but must sometimes be complemented by constitutional rules and principles.

6. REFERENCES

AUTORITEIT PERSOONSGEGEVENS, "Dutch DPA imposes a fine of 290 million euro on Uber because of transfers of drivers' data to the US", 26 August 2024, https://www.autoriteitpersoonsgegevens.nl/en/current/dutch-dpa-imposes-a-fine-of-290-million-euro-on-uber-because-of-transfers-of-drivers-data-to-the-us.

AUTORITEIT PERSOONSGEGEVENS, "AP: scraping bijna altijd illegaal", 1 May 2024, https://autoriteitpersoonsgegevens.nl/actueel/ap-scraping-bijna-altijd-illegaal.

AUTORITE DE LA CONCURRENCE, "Avis 24-A-05 du 28 juin 2024 relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative", https://www.autoritedelaconcurrence.fr/sites/default/files/integral_texts/2024-06/avisIA.pdf.

rules applicable to ChatGPT. Its mandate included the exchange of information, the coordination of external communication by different data protection authorities in their enforcement activities, and the identification of issues for which a common approach is needed in the context of their different enforcement actions regarding ChatGPT. This taskforce published a report in May 2024, which also mentioned the significant importance of providing further guidance on the interactions between the GDPR and the AI Act (EDPB, "Report of the work undertaken by the ChatGPT Taskforce", 23 May 2024, https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf). However, because the investigations mentioned above are still ongoing, the report only contains preliminary views on certain aspects of the cases.

67 Conseil constitutionnel, décision n°2020-834 QPC du 3 avril 2020, ECLI:FR:CC:2020:2020.834.QPC.

BRADFORD, Anu: *The Brussels Effect: How the European Union Rules the World*, New York: Oxford University Press, 2020.

CNIL, "Cahier air2023. IA et libre-arbitre: sommes-nous des moutons numériques?", 22 April 2024, https://www.cnil.fr/sites/cnil/files/2024-04/cahier_air2023.pdf.

CNIL, "Intelligence artificielle: le plan d'action de la CNIL", 16 May 2023, https://www.cnil.fr/fr/intelligence-artificielle-le-plan-daction-de-la-cnil.

CNIL, "'Bac à sable' intelligence artificielle et services publics: la CNIL accompagne 8 projets innovants", 22 November 2023, https://www.cnil.fr/fr/bac-sable-intelligence-artificielle-et-services-publics-la-cnil-accompagne-8-projets-innovants.

CNIL, "Refuser les cookies doit être aussi simple qu'accepter: bilan de la deuxième campagne de mises en demeure et actions à venir", 14 September 2021, https://www.cnil.fr/fr/refuser-les-cookies-doit-etre-aussi-simple-quaccepter-bilan-de-la-deuxieme-campagne-de-mises-en.

CNIL, "Rapport d'activité 2018", 15 April 2019, https://www.cnil.fr/sites/cnil/files/atoms/files/cnil-39e_rapport_annuel_2018.pdf.

CNIL, "Concertation citoyenne sur les enjeux éthiques liés à la place des algorithmes dans notre vie quotidienne: synthèse de la journée", 17 October 2017, https://www.cnil.fr/sites/cnil/files/atoms/files/cr_concertation_citoyenne_algorithmes.pdf.

COFONE, Ignacio; ROBERTSON, Adriana, "Consumer Privacy in a Behavioral World", *Hastings Law Journal* (2018) 69(6), pp. 1471-1508.

CONSEIL D'ÉTAT, "S'engager dans l'intelligence artificielle pour un meilleur service public", 30 August 2022, https://www.conseil-etat.fr/actualites/s-engager-dans-l-intelligence-artificielle-pour-un-meilleur-service-public.

EDPB, "Report of the work undertaken by the ChatGPT Taskforce", 23 May 2024, https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf.

EDPB-EDPS, "Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)", 18 June 2021, https://www.edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf.

EU AGENCY FOR FUNDAMENTAL RIGHTS, "Lack of resources undermine EU data protection enforcement", 11 June 2024, https://fra.europa.eu/en/news/2024/lack-resources-undermine-eu-data-protection-enforcement.

EUROPEAN COMMISSION, "Speech by EVP Margrethe Vestager at the European Commission workshop on 'Competition in Virtual Worlds and Generative AI'", 28 June 2024, https://ec.europa.eu/commission/presscorner/detail/en/speech_24_3550.

EUROPEAN COMMISSION, 2030 Digital Compass: the European way for the Digital Decade, 9.3.2021, COM(2021) 118 final.

EUROPEAN COMMISSION, White Paper on Artificial Intelligence - A European approach to excellence and trust, 19.2.2020, COM(2020) 65 final.

EUROPEAN COMMISSION, *Shaping Europe's Digital Future*, Luxembourg: Publications Office of the European Union, 2020, https://commission.europa.eu/system/files/2020-02/communication-shaping-europes-digital-future-feb2020_en_4.pdf.

EUROPEAN COMMISSION: LEGAL SERVICE, *70 years of EU law: a Union for its citizens*, Luxembourg: Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2880/02622.

Bottom of Form

HACKER, Philipp; CORDES, Johann; ROCHON, Janina, "Regulating Gatekeeper Artificial Intelligence and Data: Transparency, Access and Fairness under the Digital Markets Act, the General Data Protection Regulation and Beyond", *European Journal of Risk Regulation* (2023) 15(1), pp. 49-86.

MARGONI, Thomas; KRETSCHMER, Martin, "A deeper look into the EU text and data mining exceptions: Harmonisation, data ownership, and the future of technology", *GRUR International* (2022) 71(8), pp. 685-701.

META, "Building AI Technology for Europeans in a Transparent and Responsible Way", 10 June 2024, https://about.fb.com/news/2024/06/building-ai-technology-for-europeans-in-a-transparent-and-responsible-way/.

MILMO, Dan, "ChatGPT reaches 100 million users two months after launch", *The Guardian*, 2 February 2023, https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.

NOYB, "ChatGPT provides false information about people, and OpenAI can't correct it", 29 April 2024, https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it.

NOUWENS, Midas; LICCARDI, Ilaria; VEALE, Michael; KARGER, David; KAGAL, Lalana, "Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence", *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1-13.

OECD, "AI, data governance and privacy: Synergies and areas of international co-operation", *OECD Artificial Intelligence Papers*, No. 22, Paris: OECD Publishing, 2024, https://doi.org/10.1787/2476b1a4-en.

Capítulo 5

ANALYSING THE INTERPLAY BETWEEN DATA SPACES AND ARTICLE 10 OF THE AI ACT: A CASE STUDY OF CREDITWORTHINESS AI SYSTEMS

Andrés Chomczyk Penedo

Affiliated researcher. Law, Science, Technology and Society research group. Vrije Universiteit Brussel.

Anna Capellà i Ricart

Postdoctoral researcher. Institute of Law and Technology (IDT). Universitat Autònoma de Barcelona.

ABSTRACT: This paper explores the challenges Article 10(5) of the AI Act faces in ensuring that special categories of personal data can be used to mitigate biases in AI datasets. Article 10(5) imposes strict limitations on processing such data, particularly regarding data sharing. Simultaneously, the EU is promoting data spaces, i.e., common ecosystems for data sharing to facilitate data-driven innovation. This presents a potential clash between regulatory objectives. As such, this paper provides a comprehensive understanding of the legal challenges in balancing the need to access personal data to develop unbiased AI systems while limiting access to such data in a context where data is expected to flow freely. Focusing on financial services, the paper examines creditworthiness AI systems as a case study. These systems, labelled high-risk and therefore subject to Article 10(5), can benefit from specific data sharing rules within the financial data space, also known as open finance. Through this case study, the paper illustrates the complex interplay between regulatory requirements and the practicalities of data sharing in high-risk AI systems, offering insights and recommendations for policymakers and stakeholders.

KEYWORDS: data spaces, algorithmic bias, discrimination, financial services, creditworthiness

1. INTRODUCTION

The development of the EU digital economy has been at the centre of several policy strategies in the last decades (Mariniello, 2022). In this respect, a wide range of technological developments have caught the attention of EU policy and lawmakers; however, there is one topic that has received further interest: data. Currently, the main policy document dealing with it, from which different regulatory actions emerge, is the EU 2020 Data Strategy (European Commission, 2020).

At the same time, the relevance of artificial intelligence (AI) developments has also taken ground in the vision of the EU digital economy.

On the one hand, the EU is actively promoting the concept of data spaces—shared ecosystems designed to facilitate secure and efficient data sharing and collaboration under its EU 2020 Data Strategy, alongside the development of different regulatory instruments to ensure the free flow of data across the EU (Chomczyk Penedo, 2024). To facilitate data-driven innovation, these data spaces would enable the sharing of (personal) data between different stakeholders, as long as compliance with EU data protection laws is followed (Curry, Scerri & Tuikka, 2022). By developing these, the EU intends to consolidate a common single market for data across its Member States. However, while these data spaces would enable the sharing of data, they also introduce restrictions on how data can flow between different parties, for example one of the regulatory proposals that integrate the European Financial Data Space, the Financial Data Access Regulation (FiDAR proposal).[1]

On the other hand, to tackle a wide range of potential harms and risks that the use of AI can produce, the EU has adopted the AI Act, which represents a groundbreaking regulatory step. From the wide range of potential issues that could compromise fundamental rights, the existence of biases within these systems has demonstrated a serious threat (Laupman, Schippers & Papaléo Gagliardi, 2022).

In this respect, certain personal data categories, such as racial or ethnic origin or trade union membership, have long been excluded as characteristics to consider in decision-making, given their potential to expose individuals to further discrimination (Kelly et al, 2022). However, research in algorithmic discrimination field has shown that the processing of these categories can help to overcome existing biases in AI systems (Hoffmann *et al.*, 2022). As such, Article 10(5) of the AI Act specifically addresses the use of special categories of personal data to detect and correct biases.

1 Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a framework for Financial Data Access and amending Regulations (EU) No 1093/2010, (EU) No 1094/2010, (EU) No 1095/2010 and (EU) 2022/2554 COM/2023/360 final. This proposal should be read alongside the update to the current open banking framework under the Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on payment services in the internal market and amending Regulation (EU) No 1093/2010 COM/2023/367 final.

While there are plenty of regulatory crossroads and tensions at this stage of the EU digital economy, an under-researched area lays in the intersection between the AI Act and the data spaces sectoral regulation.[2] This interplay is quite relevant since Recital 68 AI Act would hint at the possibility of using these data spaces to ensure AI developers a way to access high quality data sets (as the recital points out: "*European common data spaces (…) will be instrumental to provide trustful, accountable and non-discriminatory access to high-quality data for the training, validation and testing of AI systems*"). In this respect, our paper aims to provide a preliminary analysis of Article 10(5) AI Act and how it should be read in conjunction with data spaces proposed regulations. Given the wide range of AI applications and data spaces, our study shall rely on a case study around consumer creditworthiness AI systems, which are categorised by the AI Act as high-risk. [3]

The paper is structured as follows. First, we give an overview of the AI Act, with a particular focus on its Article 10(5) (section 2). Next, we discuss the notion and purpose of data spaces as the selected instrument to enable data sharing (section 3). Then, we layout the basis of our case study around creditworthiness AI systems in financial services (section 4), alongside some identified potential clashes between Article 10(5) and the current FiDAR proposal. Based on this, and to ensure the actual utility of Article 10(5), we provide some recommendations and conclusions to address the apparent limitations found in data spaces regulatory proposals in the context of the case study (section 5).

2. OVERVIEW OF ARTICLE 10(5) OF THE AI ACT

After almost three years since its proposal, the final text of AI Act has been published in the Official Journal of the European Union.[4] The final text reveals a regulatory instrument that its main purpose is to improve the functioning of the internal market, promote the uptake of human-centric and trustworthy artificial intelligence and ensure a high level of protection of health, safety and fundamental rights.[5]

It classifies AI systems according to the risks they pose to health, safety and fundamental rights. Those that are classified as high-risk must comply with the requirements established in

2 The only exception is that of healthcare where extensive research has been conducted regarding the impact of the EDHS proposal, as for example based on de Zegher *et al.* (2024), Biasin, Yasar & Kamenjasevic (2023) or Kiseleva and de Hert (2020).

3 Annex III.5.B AI Act: *"AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score (…)".*

4 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance PE/24/2024/REV/1 OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj.

5 Article 1 AI Act.

Articles 9 to 15. In this context, Article 10 is dedicated to data and data governance requirements for high-risk AI systems.

One of the concerns behind the use of AI is its potential discriminatory effect. In this regard, the literature has warned that machine learning techniques can deduce special categories of data from innocuous proxy variables, rejecting the idea of fairness through unawareness (Williams *et al.,* 2018; Hoffmann *et al.*, 2022). Consequently, data plays a crucial part in ensuring that AI systems do not cause such situations. As such, it has been highlighted that to effectively avoid discriminatory outputs in machine learning it is relevant that protected attributes are available during model training and evaluation to account for subtle correlations and test and optimise fairness metrics (Deck *et al.*, 2024).

The protection against discrimination is explicitly included in several recitals of the AI Act, e. g. 31, 56, 57, 58, 59 or 67. Given the conflict regarding fairness through unawareness and the potential to cause discrimination of AI systems, Article 10(5) AI Act allows for the processing of personal data that constitutes special categories of personal data under GDPR.[6]

However, Article 10(5) does not provide a *carte blanche* for the processing of these special categories of data but rather it establishes certain conditions for this to occur. First, it requires that the detection and correction of bias cannot be achieved in any other way; in this sense, it requires a concrete necessity for such data. In second place, the data in question are subject to limitation on its re-use and to the state of art security and privacy-preserving measures. Thirdly, data are subject to measures to ensure that is secured, protected, subject to suitable safeguards, including strict controls and access record keeping. Fourth, data are not to be transmitted, transferred or otherwise accessed by other parties besides the entity addressing the bias in the AI system. As a fifth requirement, data used is deleted after bias detection and correction or at the end of its retention period. And finally, in sixth place, the records of processing activities, as required under GDPR, include a justification of the necessity of such processing and why it was not possible to achieve the objective by processing other data.

One of the main questions that can emerge from this preliminary analysis of Article 10(5) is where AI developers will get access to these data to address the possible existence and, if so, the tackling of the identified bias. While the Article does not provide an answer to this question, Recital 68 would hint at the use of data available through data spaces. However, Recital 68 merely addresses the access to such data and does not answer how to comply with the strict requirements set by Article 10(5).

In this respect, we can identify some questions. Firstly, if information in a data space is used to mitigate biases, does this limit the possibility of transmitting, transferring or otherwise provide access to other parties as listed in Article 10(5) to the data in the data space? Secondly, if data from the data spaces is used, the deletion obligation affects only the dataset created to

6 Since it goes beyond the scope of the paper, we will merely mention that this provision has been discussed in the literature with regards to whether it creates a new legal basis for the processing of special categories of personal data beyond those contemplated under Article 9(2) GDPR (van Bekkum and Zuiderveen Borgesius, 2022). As an argument against the creation of a new exception, Recital 70 of the AI Act adds that this processing can be done'(…) as a matter of substantial public interest within the meaning of Art. 9(2)(g) [GDPR]'.

develop the (to the best extent possible) debiased AI system or affects also the data space? As such, there are some potential tensions regarding the interaction of this provision with the envisaged data spaces. To answer these questions, we need to understand what data spaces are and how can they help in this challenge of addressing biases in AI systems.

3. THE CONCEPT AND ROLE OF DATA SPACES

The European Commission has defined data spaces differently across policy documents and regulatory proposals (Chomczyk Penedo, 2024). In general, it can be argued that a data space is an interoperable environment where data can be shared and accessed under well-defined conditions, ensuring trust, security, and data sovereignty.

The EU has been at the forefront of promoting data spaces through various initiatives and regulatory frameworks. The EU 2020 Data Strategy outlines the vision for a single European data space, where data can flow freely across sectors and borders, fostering innovation and economic growth. This strategy emphasizes the creation of nine sector-specific data spaces: industrial (manufacturing), Green Deal, mobility, health, finance, energy, agriculture, public administration, skills.

These are pivotal in unlocking the potential of big data and AI, as they enable the pooling of diverse datasets, which might lead to more accurate and unbiased AI models. This can drive advancements in AI, as diverse and high-quality datasets are essential for training robust AI models (Trigo Kramcsák, 2023).

In consequence, it is interesting to explore how data spaces can help in the application of Article 10(5) AI Act. While Recital 68 used as an example the European Health Data Space, there are eight other fields where the balancing needs to take place. For this purpose, we can look into the only other field where a regulatory proposal for a data space exists: financial services.

4. CASE STUDY: CREDITWORTHINESS AI SYSTEMS IN FINANCIAL SERVICES

Given the wide range of applications that AI can have, as well as the varied data spaces that are intended, we will rely on a case study to highlight some of the tensions that might emerge. For this purpose, we will present the case of algorithmic credit scoring systems and their framing in the AI Act (4.1). Then, we will analyse how these systems are included in the sector specific data space, the FiDAR proposal (4.2). Finally, we will present the interplay between data spaces and detection and correction of algorithmic biases in this scenario (4.3).

4.1. Algorithmic credit scoring systems and their discriminatory outcomes

In the financial services sector, creditworthiness AI systems play a crucial role in assessing the credit risk of individuals and businesses to evaluate the likelihood of a borrower defaulting on a loan (Aggarwal, 2021). These systems analyse a wide range of data, including financial history, employment status, and other personal information, to generate credit scores and risk assessments (Langenbucher, 2020). The accuracy and fairness of these assessments are critical, as they directly impact individuals' access to financial services and credit. In this sense, creditworthiness is a field that has attracted a lot of attention concerning claims of algorithmic discrimination (Bellucci *et al.*, 2010; Clarke and Rothenberg, 2018; Bicharra Garcia, 2023).

In the EU context, the recent Consumer Credit Directive[7] excludes certain data categories from the information that can be used to assess the creditworthiness of an individual, including special categories of personal data, as detailed in its Article 18(3). The exclusion of special categories of personal data from the information that can be used to assess the creditworthiness of an individual has the aim to prevent discriminatory outcomes, as it was pointed out in the decision of the Finish National Non-Discrimination and Equality Tribunal, that concluded that an algorithmic credit scoring system that based the decision of refusal of granting a credit on the combination of characters such as gender, age and place of residence was discriminatory.[8]

Although special categories of personal data cannot be used during the deployment of these AI systems, they might play a fundamental role in its training phase to detect and correct biases. Otherwise, we risk creating an AI system based around an individual that is not fully representative of the current societal landscape. As noted by Langenbucher (2022), before the deployment of these AI systems, it is necessary to train them; by doing so, AI developers create a picture of who is and who is not creditworthy in a given moment. If the societal configuration that supported that model changes, it is only reasonable to expect that it also is modified. If not, the AI systems might be biased towards certain group of individuals whose situation has changed. For example, it has been highlighted that women had been discriminated against as AI systems were trained on an ideal debtor who was a man (Kelley et al., 2023). As such, to avoid having the AI system treating more unfavourably certain people, it is necessary to debias it, even if that implies processing special categories of personal data.

4.2. Algorithmic credit scoring systems and the FiDAR proposal

One of the two regulatory instruments that will compose the European Financial Data Space is the FiDAR (now in the proposal stage). Expanding on the success of open banking,

7 Directive (EU) 2023/2225 of the European Parliament and of the Council of 18 October 2023 on credit agreements for consumers and repealing Directive 2008/48/EC PE/22/2023/REV/1 OJ L, 2023/2225, 30.10.2023.

8 See, for a summary of the decision: https://www.yvtltk.fi/en/index/opinionsanddecisions/decisions.html#

the FiDAR proposal seeks to create a more integrated and competitive financial services market by enhancing data sharing and accessibility among the different stakeholders, leading to the consolidation of an open finance framework.

The FiDAR proposal intends to establish rules on the access, sharing and use of certain categories of customer data in financial services, as detailed in its Article 1. It covers a selected group of data and financial institutions that will be engaged in this data space. While dealing with financial data, the proposal does not define this concept expressly but rather deals with 'customer data', that is defined as personal and non-personal data that is collected, stored and otherwise processed by a financial institution as part of their normal course of business, under Article 3(3). Financial data could, in certain scenarios be considered as a special category of personal data (Chomczyk Penedo and Trigo Kramcsak, 2023), as it can reveal information specially protected, for example if someone donates money to a political party (revealing their political opinions) or pays a monthly contribution to a trade union.

When it comes to creditworthiness AI systems and the data related to them, the FiDAR proposal adopts a particular approach. In this respect, we can distinguish between the data used to train the AI system, the data used to operate it, and the data resulting from it. In this respect, the FiDAR proposal only includes within its scope data about the creditworthiness, i.e., the data used to operate the AI system and obtain a credit score; it then furthers limits by only focusing on firms and not consumers. When it comes to the data related to consumers and their creditworthiness, Recital 18 highlights that the risk of exclusion outweighs the benefits from sharing data related to consumers. As for the other data categories included in the scope of the FiDAR proposal, its use for credit scoring activities shall be subject to limitations to be established by the European Banking Authority and the European Data Protection Board through the data use perimeters, as provided for under Article 7(2).

While these provisions tackle the use of data by AI creditworthiness systems and the possibilities of getting access to other data categories for the conduction of a creditworthiness assessment, these do not answer how to deal with the training of the AI system itself. Moreover, it also does not tackle whether certain data, that can be considered as a special category of personal data, generated by a financial data space participant can be made available in the data space for others to improve their systems.

4.3. Interplay and clashes between data spaces and detection and correction of algorithmic biases

In Article 10(5) AI Act, we find no limitations regarding the origin of the data. In this respect, it could be data already collected by, based on our case study, a bank, for example; also, it could be obtained through the data space if uploaded by another participant, for example a credit bureau. Here we understand that all persons that access the data space are only authorised persons with appropriate confidentiality obligations (Article 10(5)(c)) and that data gathered from the data space afterwards cannot be accessed by other parties even

though in the data space will still be available (Article 10(5)(d)). In this respect, the relevant financial data sharing scheme and data use perimeter, as provided for under Articles 7 and 10 of the FiDAR proposal respectively, would lay out the applicable confidentiality obligations required by the AI Act.

Furthermore, Article 10(5) requires the erasure of the data used, after it has served its purpose. If the data was obtained via a data space, we can ask ourselves: complying with Article 10(5)(e) implies deleting the data only from the dataset the AI developer used or also from the data space itself? In this respect, we can propose that, for Article 10(5) AI Act, it is only required the deletion from the AI developer side after the task is completed.

It must be acknowledged the risk that the data included in these spaces used to train algorithmic credit scoring systems may lead to biased results if there is no process of bias detection or correction (Balayn and Gürses, 2021; Van Bekkum and Zuiderveen, 2023).

5. CONCLUDING REMARKS AND POSSIBLE RECOMMENDATIONS

As analysed in our contribution, the role of data spaces as sources for the necessary data to de-bias an AI system might not be as straightforward as envisaged by the AI Act. This is because some of these necessary data categories might not be available under the specific sectoral data space regulation; and, even when available, these might be subject to further requirements. Recital 70 of the AI Act justifies the processing of special categories of data to correct and mitigate bias as a substantial public interest. In this respect, we can argue whether the answers to balance between these different legal instruments lays elsewhere. Given that we are dealing with personal data, GDPR inevitably comes to the foreground in this process through its principles.

In this respect, the principle of data minimisation could offer some answers. Under the FiDAR proposal, data use perimeters should be established,[9] meaning that personal data shall be limited to what is necessary concerning the purposes for which they are processed, despite the intention of ensuring a free flow of data. At the same time, Article 10(5) AI Act also intends to limit data used to bias detection and correction as the minimum possible. However, it is possible to question if, to effectively perform a mandated task, maybe the more relevant data available the better, even if that means involving data categories originally excluded. From our perspective, data spaces can be used to gather a quantity and diversity of data -only available there- that can be very useful to correct and detect biases of AI systems, such as creditworthiness systems.

In this sense, for example, the FiDAR proposal has the necessary tools to ensure that special categories of personal data are adequately used to detect and correct biases. In this regard, we suggest that when drafting the guidelines for the data use perimeters as well as the financial data sharing schemes, the objectives behind provisions such as Article 10(5) AI Act are

9 Article 7 FiDAR proposal

taken into consideration. In our case study, this means that the role of the European Banking Authority as well as of the European Data Protection Board might prove to be crucial to harmonise these two regulatory frameworks, establishing guidelines on the process to obtain and use the data from the data space and controlling that the data is adequately utilised.

We consider that the benefits emerging from allowing the use of such data can outweigh the risks from not allowing its use considering that if nothing is done, creditworthiness assessment AI systems will continue to present avoidable discriminatory outcomes. It could be argued that enabling the use of creditworthiness AI systems without granting access to data that could be used to tackle bias (implementing always the proper safeguards) fosters a consolidation of further financial exclusion rather than a promotion of financial inclusion.

6. REFERENCES

AGGARWAL, Nikita: "The norms of algorithmic credit scoring." The Cambridge Law Journal (2021) 80(1), pp. 42–73 https://doi.org/10.1017/S0008197321000015

BALAYN, Agathe, and GÜRSES, Seda: "Beyond debiasing. Regulating AI and Its Inequalities". European Digital Rights (2021). https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

BELLUCCI, Andrea *et al.*: "Does gender matter in bank–firm relationships? Evidence from small business lending". *J Bank Financ.* (2010) 34(12), pp. 2968–2984 https://doi.org/10.1016/j.jbankfin.2010.07.008

BICHARRA GARCIA, Ana Cristina: "Algorithmic discrimination in the credit domain: what do we know about it?", *AI & Soc* (2023). https://doi.org/10.1007/s00146-023-01676-3

BIASIN, Elisabetta., YASAR, Burcu, and KAMENJASEVIC, Erik: "New cybersecurity requirements for medical devices in the EU: the forthcoming European health data space, data act, and artificial intelligence act". *Law, Tech. & Hum.* (2023) 5(2), pp. 43-58. https://doi.org/10(5)204/lthj.3068

CHOMCZYK PENEDO, Andrés: "The Regulation of Data Spaces under the EU Data Strategy: Towards the 'Act-ification' of the Fifth European Freedom for Data?" European Journal of Law and Technology (2024) 15(1), pp. 1-26. https://www.ejlt.org/index.php/ejlt/article/view/995/1088

CHOMCZYK PENEDO, Andrés and TRIGO KRAMCSÁK, Pablo: Can the European Financial Data Space remove bias in financial AI development? Opportunities and regulatory challenges, *International Journal of Law and Information Technology* (2023) 31(3), pp. 253–275, https://doi.org/10.1093/ijlit/eaad020

CLARKE, Kevin A., and ROTHENBERG Lawrence S.: "Mortgage pricing and race: evidence from the northeast". *Am Law Econ Rev.* (2018) 20(1), pp. 138–167. https://doi.org/10.1093/aler/ahx021

CURRY, Edward, SIMON, Scerri, and TUOMO, Tuikka: *Data Spaces: Design, Deployment and Future Directions*. Springer Nature, 2022.

DE ZEGHER, Isabelle *et al.*: "Artificial intelligence based data curation: enabling a patient-centric European health data space". *Front. Med.* (2024) 11:1365501, pp. 1-12. 10.3389/fmed.2024.1365501

DECK, Luca *et al.*: "Implications of the ai act for non-discrimination law and algorithmic fairness", *ArXiv* (2024), pp. 1-7. https://arxiv.org/pdf/2403.20089

EUROPEAN COMMISSION: COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A European strategy for data COM/2020/66 final, Brussels, 2020

HOFFMANN, Hanna, *et al.*: "Fairness by awareness? On the inclusion of protected features in algorithmic decisions", *Computer, Law and Security Review* (2022) 44, pp. 1-12.

KELLEY, Stephanie, OVCHINNIKOV, Anton, HARDOON, David R., and HEINRICH, Adrienne: "Antidiscrimination laws, artificial intelligence, and gender bias: A case study in nonmortgage fintech lending". *Manufacturing & Service Operations Management* (2022) *24*(6), pp. 3039-3059.

KISELEVA, Anastasiya, and DE HERT, Paul: "Creating a European Health Data Space: obstacles in four key legal area". *EPLR* (2021) 5(1), pp. 21-36. https://doi.org/10.21552/eplr/2021/1/5

TRIGO KRAMCSÁK, Pablo: "Can legitimate interest be an appropriate lawful basis for processing Artificial Intelligence training datasets?". *Computer Law & Security Review* (2023) *48*, 105765. 10.1016/j.clsr.2022.105765

LANGENBUCHER, Katja: "Responsible A.I.-based credit scoring – a legal framework', European Business Law Review, (2020) 31(4), pp. 527-572. https://doi.org/10(5)4648/eulr2020022

LAUPMAN, Clarisse, and SCHIPPERS, Laurianne-Marie, PAPALÉO GAGLIARDI, Marilia: "Biased Algorithms and the Discrimination upon Immigration Policy", a: Custers, Bart and Fosch-Villaronga, Eduard (eds*): Law and Artificial Intelligence. Information Technology and Law Series, vol 35,* The Hague: T.M.C. Asser Press, 2022, pp. 187-204. https://doi.org/10.1007/978-94-6265-523-2_10ob

VAN BEKKUM, Marvin and ZUIDERVEEN, Frederik: Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law and Security Review* (2023) 48, 105770. https://doi.org/10.1016/j.clsr.2022.105770

WILLIAMS, Betsy Anne *et al.*: "How Algorithms Discriminate Based on Data they Lack: Challenges, Solutions, and Policy Implications*", Journal of Information Policy* (2018) 8, pp. 78-115.

CAPÍTULO 6

GENERATIVE AI CONTENT MISUSE AND THE DSA

Ioannis REVOLIDIS

Resident Academic Lecturer
University of Malta, Faculty of Laws, Centre for Distributed Ledger Technologies

ABSTRACT: This paper explores the challenges posed by Generative AI applications, particularly their potential to produce fake news, misinformation, and disinformation, thereby complicating content moderation efforts. The study delves into the applicability of the Digital Services Act (DSA) to Generative AI tools, presenting a spectrum of opinions from legal commentators. While some argue that the DSA's scope is too narrow, others advocate for more flexible interpretations, suggesting classifications such as search engines or hosting providers. A common consensus emerges that Generative AI can be regulated under the DSA in certain contexts, particularly when integrated into traditional content-sharing platforms. The paper critically assesses the opinions expressed in legal literature so far and identifies potential gaps. It also highlights the necessity for further research to identify the most suitable intermediary framework within the DSA and to develop an effective mix of liability, potential immunity, and content moderation framework.

KEYWORDS: Generative AI, Digital Services Act (DSA), Content Moderation, Misinformation, AI Regulation

1. INTRODUCTION.

The advent of deepfakes[1] and AI-generated media constitutes another chapter[2] in the long history of content misuse and manipulation[3]. Their existence captured the collective imagination almost instantaneously and swiftly took a sinister turn[4]. It was becoming increasingly evident that, alongside traditional media manipulation tools, artificial intelligence was entering the mainstream. Its broader availability promised to exacerbate the production of content further. Such production does not necessarily possess an illegal or harmful character in every instance[5]. Nevertheless, considering that the accessibility of these tools can invariably function as a double-edged sword, each significant milestone in the development and utilisation of AI-generated media inherently carries the potential to broaden the operational scope of malicious actors. In other words, the dual nature of these advancements means that while they can be harnessed for innovative and beneficial purposes, they simultaneously open new avenues for misuse and exploitation by those with malicious intent.

1 On deepfakes see, among many others, WESTERLUND, Mika, "The Emergence of Deepfake Technology: A Review", Technology Innovation Management Review, (2019) 9(11), pp. 39-52.

2 For an insightful breakdown of classic mass media manipulation frameworks see HERMAN, Edward; CHOMSKY, Noam: Manufacturing Consent - The Political Economy of the Mass Media, digital edition, London: The Bodley Head Random House, 2008.

3 For a historical perspective on media manipulation and fake news see SAFIEDDINE, Fadi: "History of Fake News", in: Ibrahim, Yasmin; Safieddine, Fadi (eds.): Fake News in an Era of Social Media: Tracking Viral Contagion, London: Rowman&Littlefield, 2020, pp. 1-26, GORBACH, Julien, "Not Your Grandpa's Hoax: A Comparative History of Fake News", American Journalism, (2018) 35(2), pp. 236–249.

4 By 2017, widely regarded as the inaugural year of deepfake culture, numerous instances of deepfake pornography—typically involving AI-generated face swaps—had begun to populate various websites across the internet. For an assessment of the phenomenon see REISSMAN, Hailey, "What Is Deepfake Porn and Why Is It Thriving in the Age of AI?", available at https://www.asc.upenn.edu/news-events/news/what-deepfake-porn-and-why-it-thriving-age-ai (last access 30.07.2024).

5 On the contrary, Generative AI output can also be deployed for beneficial purposes, see, for example, LEES, Dominic, "Deepfakes are being used for good – here's how", available at https://research.reading.ac.uk/research-blog/deepfakes-are-being-used-for-good-heres-how/ (last access 30.07.2024).

Indeed, over the subsequent years, the proliferation of AI-generated content has expanded significantly[6], both in terms of the diversity of their applications[7] and in terms of the sophistication[8] of the techniques employed[9].

It remains evident that malicious actors could employ the same techniques to disseminate fake, misleading, or outright false information to the public. Such incidents have already been recorded. For example, in the initial weeks following the outbreak of the conflict between Russia and Ukraine, a deepfake video surfaced featuring Ukraine's President, Volodymyr Zelenskyy[10]. In this manipulated footage, he appeared to concede that Ukraine had been defeated and urged Ukrainian soldiers to lay down their weapons. More recently, amidst an increasingly polarised political climate in the West, various social media platforms have been inundated with AI-generated photographs that purportedly depicted the dramatic pursuit and arrest of former President and current presidential candidate Donald Trump[11].

6 For an overview and a timeline of the early period of AI generated content see Homeland Security, "Increasing Threats of Deepfake Identities", available at https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf (last access 30.07.2024).

7 For instance, in 2018, Jordan Peele, in collaboration with Buzzfeed, utilised the likeness of former President of the United States, Barack Obama, to create a public service announcement (PSA) aimed at highlighting the issue of untrustworthy news sources online. This was a particularly timely experiment given the emerging threat posed by deepfakes. Peele's video was an exemplary demonstration of the so-called lip sync technique, which involves several steps to materialise. At the conclusion of his experiment, Jordan Peele revealed his own face while manipulating the likeness of Barack Obama. This move was intended to demystify the process and to alert internet users to this potential threat. For an overview of Jordan Peele's deep fake project see SILVERMAN, Craig, "How To Spot A Deepfake Like The Barack Obama–Jordan Peele Video", available at https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed (last access 30.07.2024) and ROMANO, Aja, "Jordan Peele's simulated Obama PSA is a double-edged warning against fake news", available at https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed (last access 30.07.2024).

8 Another example utilising a more sophisticated technology is the deep fake video of Richard Nixon making a fake moon landing speech. Beginning in 2019, media artists Francesca Panetta and Halsey Burgund, affiliated with the Massachusetts Institute of Technology, embarked on a collaborative project with two artificial intelligence companies, Canny AI and Respeecher. Their ambitious endeavour aimed to create a posthumous deepfake video. The resulting synthetic footage portrays former President Richard Nixon delivering a speech that he had never intended to give, a full fifty years after the Apollo 11 mission. In this case, the creators deployed the more sophisticated technique called "puppet deepfake" or "puppet master", a process by virtue of which the head movement and facial expressions are transferred in real-time to the liking of another individual. For a more detailed outlook on this incident see DELVISCIO, Jeffery, "A Nixon Deepfake, a 'Moon Disaster' Speech and an Information Ecosystem at Risk", available at https://www.scientificamerican.com/video/a-nixon-deepfake-a-moon-disaster-speech-and-an-information-ecosystem-at-risk1/ (last access 30.07.2024).

9 For a detailed analysis about deep fake technology, techniques, potential use cases and detection see FARID, Hany, "Creating, Using, Misusing, and Detecting Deep Fakes", Journal of Online Trust and Safety, (2022) 1(4), pp. 1-33.

10 For this particular incident but also for the deployment of deep fakes in warfare more general see TWOMEY, John Joseph; LINEHAN, Conor; MURPHY, Gillian, "Deepfakes in warfare: new concerns emerge from their use around the Russian invasion of Ukraine", available at https://theconversation.com/deepfakes-in-warfare-new-concerns-emerge-from-their-use-around-the-russian-invasion-of-ukraine-216393 (last access 30.07.2024).

11 For the wider dimensions of this incident see GARBER, Megan, "The Trump AI Deepfakes Had an Unintended Side Effect", available at https://www.theatlantic.com/culture/archive/2023/03/fake-trump-arrest-images-ai-generated-deepfakes/673510/ (last access 30.07.2024).

While the examples mentioned above primarily involve AI manipulation that comes from organised experts and necessitates a certain level of technical expertise and access to substantial resources, the increasing availability of Generative AI applications to a broader spectrum of internet users significantly lowers the barrier to entry. This expanding accessibility paves the way for a more extensive deployment of synthetic content by individuals who may lack considerable technical skills or resources. Consequently, the accessibility of these user-friendly AI tools "democratises" the creation of synthetic media, enabling a vast array of users to generate and disseminate manipulated content with relative ease. Recent research confirms that the availability of Generative AI tools and their utilisation by non-institutional average users should not be underestimated but, on the contrary, that the diffusion of productivity tools enhances the potential for the production and distribution of disinformation, fake and misleading content[12].

The problem of AI-driven content misuse is not solely related to human geographies and tool accessibility. It also concerns the functioning of democratic institutions and the way users engage with digital information. As already pointed out in literature[13], the risks posed by AI-powered misuse of content could be clustered in four major categories: a) Manipulation of elections. Deepfakes can distort democratic processes by spreading false material to influence election outcomes. Both foreign states and individuals with basic technical skills pose a risk of deploying such deepfakes, b) Exacerbate social divisions. In a politically charged environment, deepfakes can deepen social divides by presenting manipulated videos that reinforce polarising views on issues like economic inequality, race, and sexuality. This can lead to increased societal discord and, in some cases, incite harmful actions, c) Erode trust in institutions. Deepfakes can undermine trust in public institutions by depicting fake scenarios of misconduct by officials, such as police brutality or judicial corruption. This erosion of trust poses significant risks to democracy and public safety and d) Undermine journalism and user engagement with media. The growing difficulty in distinguishing real from fake content undermines trust in the media. Deepfakes increase the challenge for journalists to verify information quickly, leading to public scepticism and potentially causing news outlets to hesitate in publishing stories. Lesser-known deepfakes that don't attract immediate attention can still cause long-term harm by subtly influencing public opinion.

This new mode of engagement with digital applications and the associated risks raise several critical issues concerning content supervision, curation, moderation, and control.

12 See HASSOUN, Amelia; BORENSTEIN, Gabrielle; OSBORN, Katy; McAULIFFE, Jacob; GOLDBERG, Beth, "Sowing "seeds of doubt": Cottage industries of election and medical misinformation in Brazil and the United States", New Media & Society, (2024) 0(0) (Online First https://journals.sagepub.com/doi/10.1177/14614448241255379, last access 30.07.2024).

13 WALDEMARSSON, Christoffer, "Disinformation, Deepfakes & Democracy", available at https://www.allianceofdemocracies.org/wp-content/uploads/2020/04/Disinformation-Deepfakes-Democracy-Waldemarsson-2020.pdf (last access 30.07.2024), pp. 10-11.

Most importantly, it prompts questions about the necessity and form of a regulatory intervention. The form that such regulatory intervention takes must, however, be based on an examination of existing legal frameworks and their capacity to effectively address the problem at hand.

While the AI Act has now been published in the Official Journal of the European Union[14], it does not primarily address issues related to content. It is true that art. 50 of the EU AI Act attempts to mitigate certain aspects of synthetic content[15], but this is not a comprehensive solution to content-related challenges. Conversely, one might turn to the Digital Services Act (henceforth DSA), which encapsulates the more established EU framework with regards to user-generated content and the role of the digital infrastructure that is involved in the process of storing and disseminating it.

This paper will primarily focus on the relationship between the DSA and Generative AI tools that facilitate the production of user-generated content. Specifically, it will explore whether Generative AI applications can be regulated under the safe harbour provisions of the DSA. To achieve this objective, the paper is structured into two parts. The first part will present the current literature discussing the applicability of the DSA to Generative AI tools. The second part will critically assess the positions expressed in the current discourse and will endeavour to suggest alternative avenues of interpretation. The conclusion will summarise the findings and suggest directions for future research.

2. GENERATIVE AI AND THE DSA: A SHORT LITERATURE REVIEW.

Legal literature has already begun to address the issue of whether Generative AI applications can be included within the regulatory scope of the DSA. This should come as no surprise, given that the risk profile of Generative AI applications, as established in the introduction of this paper, aligns in many respects with the regulatory goals of the DSA, particularly regarding content moderation. This section aims to categorise and present the various opinions expressed by different commentators. The subsequent part of this paper will critically assess the opinions presented in this section.

14 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) Text with EEA relevance [2024] OJ L1689/12.

15 The EU AI Act, through art. 50, aims to regulate part of the problem of deep fakes and AI Generated content by mostly imposing transparency obligations throughout the AI value chain. For a more detailed outlook see ŁABUZ, Mateusz, "Deep fakes and the Artificial Intelligence Act—An important signal or a missed opportunity?", Policy & Internet, Online Version of Record before inclusion in an issue available at https://onlinelibrary.wiley.com/doi/full/10.1002/poi3.406 (last access 30.07.2024).

2.1. Opinions that deny the application of the DSA over Generative AI applications.

A segment of legal literature summarily rejects the notion that Generative AI applications might fall under the DSA.

For example, while Hacker, Engel, and List acknowledge that the DSA might regulate AI-generated content on traditional social media platforms, they take a firm stance against applying the DSA to Generative AI applications themselves[16]. They recognise that the risk profile of Generative AI applications raises content misuse issues similar to those of traditional digital platforms, potentially elevating misinformation, disinformation, fake news, and hate speech to unprecedented levels. Nonetheless, they argue that the DSA, in its current form, is inadequate for regulating content produced by Generative AI. Their primary argument is that Generative AI applications do not fit within any established intermediary categories under the DSA. They quickly dismiss the possibility of classifying Generative AI tools as "mere conduits" or "caching" providers and focus on whether they could be considered "hosting" providers. However, they conclude that this is not feasible, as the legislative definition in art. 3(g) of the DSA confines hosting intermediaries to storing information provided by and at the request of service recipients. They argue that the relevant content is generated by the AI applications themselves, not the service recipients[17].

In a follow-up paper, Hacker, Engel, and Mauer (replacing List) maintain their original stance, asserting that Generative AI applications cannot be regulated under the DSA as they do not fit into any intermediary categories[18]. They add that the DSA was not designed to address content produced via Generative AI applications[19]. They reiterate their core argument: Generative AI applications fall outside the DSA's scope because they do not align with the definitions of existing intermediaries. In this follow-up, they aim to reinforce their position by pointing to the CJEU decision in L' Oreal[20]. They assert that, as per the decision of the Court in L' Oreal, service providers lose DSA immunities if they "provide assistance" in terms of content management and presentation, thus leaving their neutral

16 See HACKER, Philipp; ENGEL, Andreas; LIST, Theresa, "Understanding and Regulating ChatGPT, and Other Large Generative AI Models", available at https://verfassungsblog.de/chatgpt/ (last access 30.07.2024).

17 Ibid, where they note: "... The trick with LGAIMs, however, is that the relevant content is decidedly not provided by the user, but by the LGAIM itself, having been prompted by the user via the insertion of certain query terms (e.g., "write an essay about content moderation in EU law in a lawyerly style")...".

18 See HACKER, Philipp; ENGEL, Andreas; MAUER, Marco, "Regulating ChatGPT and other Large Generative AI Models", FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, (2024), pp. 1112-1123.

19 Ibid, p. 1118.

20 Case C-324/09 L'Oréal SA and Others v eBay International AG and Others [2011] ECR I-6011, ECLI:EU:C:2011:474.

stance. They argue that Generative AI applications exceed this threshold, making them ineligible for inclusion in the DSA[21].

It is important to note that, in both publications, the authors do not deny the usefulness of the DSA's content moderation tools. In their regulatory recommendations, they essentially advocate for a DSA-style content moderation system for Generative AI applications.

In a similar vein, Lilian Edwards, Igor Szpotakowski, Gabriele Cifrodelli, Joséphine Sangaré, and James Stewart give little prospect to the application of the DSA to Generative AI applications[22]. Although the starting point of their paper is not the problem of content moderation but a very insightful analysis of the terms and conditions of Generative AI providers, they take a more holistic approach to the regulation of such applications. They are, in general, sceptical of a phenomenon they call "private ordering". With this eloquent term, they describe what they see as the effort of Generative AI providers to present themselves as neutral intermediaries, shifting all the risks of their operations to their end users via an elaborate crafting of their terms and conditions and other self-regulatory tools. As they neatly summarise their key findings "Model providers are seeking all the benefits of neutrality in terms of deferring liability and responsibility to users, while still gaining all the advantages of their position in terms of profit and power. This suggestion is bolstered by the way all or most of the providers in our sample behaved as if they were indeed platforms under the ECD (now DSA) and the DMCA in terms of content moderation; accepting DMCA notices for takedown, removing repeat infringers etc, as if this would provide them with safe harbours like any other "platform""[23]. It is in this context that they review whether the private ordering devised by Generative AI applications can find support under existing law and, more specifically, the DSA. On that front, they deny any such possibility. They assert that the DSA still aligns with the original policy framework of the Electronic Commerce Directive, which assumed platforms merely stored and shared user-generated content, granting them immunity to avoid unlimited liability for the actions of their users[24]. Initially, EU electronic commerce intermediaries were simply "messengers" facilitating information exchange and were only liable when alerted to illegal use of their services. However, like Hacker et al., they argue

21 HACKER, Philipp; ENGEL, Andreas; MAUER, Marco, "Regulating ChatGPT and other Large Generative AI Models", FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, (2024),, p. 1118: "... To the contrary, CJEU jurisprudence shows that even platforms merely storing user-generated content may easily lose their status as hosting providers, and concomitant liability privileges under the DSA (and its predecessor in this respect, the E-Commerce Directive), if they "provide assistance" and thus leave their "neutral position", which may even mean merely promoting user-generated content (CJEU, Case C-324/09, L'Oréal para 116). A fortiori, systems generating the content themselves cannot reasonably be qualified as hosting service providers. Hence, the DSA does not apply...".

22 See EDWARDS, Lilian; SZPOTAKOWSKI, Igor; CIFRODELLI, Gabriele; SANGARÉ, Joséphine; STEWART, James, "Private Ordering and Generative AI: What Can We Learn From Model Terms and Conditions?", CREATe Working Paper (2024) available at https://zenodo.org/records/11276105 (last access 30.07.2024).

23 Ibid, p. 20.

24 Ibid.

this does not apply to Generative AI tools, which create content rather than just hosting user content[25]. They question the inclusion of Generative AI under the immunity regime, noting these tools are not just passive victims of user-created content risks but are themselves content creators[26]. Despite this, they recognise that the DSA provides valuable tools for transparency and content moderation and suggest it should be amended to address Generative AI applications[27].

2.2. Intermediate positions: Generative AI tools as search engines.

Other commentators take a more intermediate approach. While they acknowledge the problem of classifying Generative AI applications under the DSA and their imperfect fit within existing intermediary categories, they search for a more flexible solution by exploring whether Generative AI tools could be regulated as search engines.

Sophie Stalla-Bourdillon views intermediary categories under the DSA as flexible concepts[28]. She uses search engines as an example, noting they also do not perfectly fit within the three basic intermediary types (mere conduit, caching, hosting) but are still regulated by the DSA. She explores the possibility of classifying Generative AI applications as hosting providers but questions whether, under art. 3(g) of the DSA, it is the users or the AI that provides the content[29]. She then focuses on search engines, drawing

25 Ibid, p. 21.

26 Ibid, p. 21: "... In policy terms, relieving a model provider of liability is inappropriate because they are not a mere hapless victim of risks deriving from user-created content, but the creator of the content themselves by allowing users to query the model …".

27 Ibid, p. 22: "... The problem then is that the DSA does not, as the ECD did, just provide liability exemptions; it also demands positive steps of hosts, platforms and VLOPs of varying natures. And these are steps that, judging by our research above, are exactly what are needed to protect the B2C users of the generative AI sector … These provisions would be extraordinarily useful in meeting the procedural vices identified above in model T&Cs and would transform their generally hostile and unfair governance approach to disempowered users. There seems no good policy reason why these rules should not be applied to foundation models. At present, model providers have their cake and eat it; they assert exemption from liability by passing risk via their terms and conditions to users, but evade the new positive obligations of the DSA. This is unjust. We suggest therefore that the DSA is already not fit for purpose and should be amended to bring foundation models within its scope as soon as possible …".

28 STALLA-BOURDILLON, Sophie, "What if ChatGPT was much more than a chatbox? What if LLM-as-a-service was a search engine?", available at https://peepbeep.blog/2023/04/03/what-if-chatgpt-was-much-more-than-a-chatbox-what-if-llm-as-a-service-was-a-search-engine/ (last access 30.07.2024).

29 Unlike Hacker et. al and Edwards et. al, nonetheless, she views this issue are more open to debate by noting "... Although the category of hosting services is probably the closest one to LLMaaS, it is not a perfect fit either. With LLMaaS, the stored information, i.e., the model output, is not [strictly speaking] provided [could provided also mean triggered?] by the recipient of the service, although its storage is performed at the request of the recipient of the service. With this said, the model input is provided by the recipient of the service and stored at the request of the recipient of the service. Would considering the model input sufficient to make LLMaaS a

functional comparisons with Generative AI applications, suggesting that such a classification might be viable.

Beatriz Botero Arcila also draws a functional comparison between search engines and Generative AI tools[30]. She acknowledges the content moderation risks posed by Generative AI and their imperfect fit within the DSA's intermediary types[31]. However, she argues that ambiguity has always been part of interpreting EU platform regulations. Botero Arcila claims that search engines and Generative AI applications share a fundamental role: organising information in an information-rich environment. This functional equivalency justifies the extension of the DSA provisions on search to Generative AI applications[32]. While recognising that Generative AI performs additional functions, she concludes that partially regulating them under the DSA's search engine provisions is fair and practicable[33].

2.3. Flexible appraisal: Generative AI tools as host providers or search engines.

Finally, in what is likely the most flexible approach, Laureline Lemoine and Mathias Vermeulen argue that Generative AI applications could be considered either host providers or search engines under the DSA[34]. Regarding hosting, they discuss whether content is provided by users or the AI, suggesting that user prompts might allow a flexible interpretation of "hosting." As for search engines, like Stalla-Bourdillon and Arcila, they highlight the ambiguities and functional similarities between search engines and Generative AI applications, concluding that Generative AI applications could fit the search engine description to a certain extent.

3. ANALYSIS.

The previous section has shown that the inclusion of Generative AI applications under the DSA's regulatory scope has elicited mixed reactions in legal literature. Opinions range

hosting service? It's probably not the best argument when the primary concern relates to the handling of the model output, but is it good enough? …".

30 BOTERO ARCILA, Beatriz, "Is it a Platform? Is it a Search Engine? It's Chat GPT! The European Liability Regime for Large Language Models", Journal of Free Speech Law, 2023 3(2), pp. 455-488.

31 Ibid, p. 477 and 479.

32 Ibid, p. 483-485.

33 Ibid, p. 486-487.

34 LEMOINE, Laureline; VERMEULEN, Mathias, "Assessing the Extent to Which Generative Artificial Intelligence (AI) Falls Within the Scope of the EU's Digital Services Act: an Initial Analysis", available at SSRN: https://ssrn.com/abstract=4702422 (last access 30.07.2024).

widely, with some commentators believing the DSA's scope is too narrow to cover Generative AI tools, while others explore more flexible models for their inclusion.

However, a common theme among commentators is that Generative AI applications raise significant content moderation issues, enabling average users to create potentially illegal, harmful, malicious, or misleading content. Another point of agreement is that AI-generated content can be regulated by the DSA at a secondary level when it appears on traditional content-sharing platforms. There is also consensus that incorporating Generative AI elements into traditional content storage and sharing applications should not be prima facie excluded from the DSA's scope.

Nevertheless, the various approaches regarding the applicability of the DSA to Generative AI applications merit individual assessment.

At one end of the spectrum, the opinion that dismisses the applicability of the DSA to Generative AI applications appears reasonable. It is fair to say that Generative AI was not the primary regulatory target of the DSA, posing challenges in finding the appropriate regulatory framework to address the risks associated with their content-related capabilities.

However, this opinion may not be entirely convincing for several reasons.

First, even if one accepts the rather reasonable assumption that the DSA did not primarily target Generative AI applications, this is not a decisive argument for their prima facie exclusion from the scope of the DSA. The realm of emerging technologies, which the DSA primarily regulates, is fast-paced and constantly evolving. This structural characteristic of digital environments justifies a flexible, creative, and dynamic interpretative approach that extends beyond the strict confines of a grammatical interpretation of the DSA's provisions[35]. This is especially true for those referring to intermediary services that might benefit from liability immunities but also face accountability obligations regarding content moderation[36]. The DSA includes mechanisms that support such a flexible approach. Recital 28 of the DSA acknowledges the evolving nature of digital applications, while Recital 29 further suggests that the classification of a specific service as a "mere conduit", "caching", or "hosting" service should depend solely on its technical functionalities, which may evolve over time and should be assessed on a case-by-case basis. In this context, it might not be unreasonable to propose that only after thoroughly exploring the outer teleological limits of the DSA in relation to Generative AI applications can a conclusion be drawn regarding their inclusion or exclusion from the regulatory framework of the DSA.

Second, the argument that Generative AI applications do not fit the notion of "hosting" providers because they produce content rather than merely storing it, is debatable and requires nuance. While it is true that Generative AI applications play a role in content pro-

35 In the same spirit, Stalla-Bourdillon, supra note 28.

36 For a more general argument on the flexibility of the intermediary phenotypes of the DSA see REVOLIDIS, Ioannis: "Internet Intermediaries and Copyright Enforcement in the EU: In Search of a Balanced Approach", in: Corrales, Marcelo; Fenwick, Mark; Forgó, Nikolaus (eds.): New Technology, Big Data and the Law, Singapore: Springer Nature, 2017, pp. 223-246, especially p. 229-232.

duction, the exact nature of this role needs careful assessment. The CJEU judgement in Papasavvas clarifies that a digital service provider falls outside the intermediary taxonomy of the DSA if they make all critical decisions about the hosted content[37]. In Papasavvas, this was evident, as the content was produced by journalists of a digital newspaper, the main defendant in a defamation case. However, Generative AI applications differ in that they are primarily tools used by individuals. These applications do not decide to create content nor influence its parameters, direction, tone, or substance; these decisions are made by the users[38]. Even if Generative AI applications significantly aid in synthesising output, content creation involves more stages than just producing the output. Thus, determining the actual creator of AI-generated content requires examining several parameters. Each use case might also need individual assessment, as the levels of human and AI involvement in content production can vary. It appears, therefore, impossible to prima facie classify all Generative AI applications as actual content creators.

Third, the argument that Generative AI applications play an active role beyond what is permissible under art. 6 of the DSA, especially in light of the findings of the CJEU in the L'Oreal case, is not entirely convincing. To begin with, L'Oreal might not be the best precedent, particularly given subsequent CJEU case law offering a more nuanced approach to content-related applications. For example, the CJEU's approach in YouTube and Cyando[39] differs from L'Oreal. In this case, the Court considered whether content storing and sharing applications could benefit from art. 14 of the Electronic Commerce Directive (the predecessor to art. 6 of the DSA). The Court did not rely on L'Oreal but instead assessed YouTube's and Cyando's business models independently. It ruled that features like automatic content indexing, search functions, and video recommendations based on user profiles did not imply that the operator had specific knowledge of illegal activities or information on the platform, something that would remove their status as passive intermediaries[40]. Had the Court relied on L'Oreal, it would have likely deemed YouTube's and

37 Case C-291/13 Papasavvas v O Fileleftheros Dimosia Etairia Ltd and Others ECLI:EU:C:2014:2209, paras 40-46.

38 HASSOUN, Amelia; ABONIZIO, Ariel; OSBORN Katy; WU, Cameron; GOLDBERG, Beth, "The Influencer Next Door: How Misinformation Creators Use GenAI", available at https://arxiv.org/abs/2405.13554#:~:text=Based%20on%20longitudinal%20ethnographic%20research,their%20personal%20needs%20and%20desires (last access 30.07.2024) demonstrate with impressive insightfulness and precision that Generative AI applications are passive tools in the hands of misinformation creators.

39 Joined Cases C-682/18 and C-683/18 YouTube and Cyando ECLI:EU:C:2021:503.

40 Joined Cases C-682/18 and C-683/18 YouTube and Cyando ECLI:EU:C:2021:503, para 114: "In that regard, the fact that the operator of an online content-sharing platform automatically indexes content uploaded to that platform, that that platform has a search function and that it recommends videos on the basis of users' profiles or preferences is not a sufficient ground for the conclusion that that operator has 'specific' knowledge of illegal activities carried out on that platform or of illegal information stored on it". Despite content storing and sharing platforms performing more functions than just storing user-generated content, the Court did not remove their intermediary status. Notably, the Court also ruled that these platforms do not perform a 'communication to the public' under art. 3(1) of the InfoSoc Directive by merely making their platform available (paras 76-102). Although this was a separate issue, the Court acknowledged (para 108) that whether these platforms

Cyando's involvement too high to remain neutral and passive. Additionally, it is important to note that while Generative AI applications appear to actively produce content, the process could be viewed as automated and passive[41], entirely dependent on user prompts. Although the content reflects usually the training of the AI model and the design choices of the developers, these choices are abstract compared to the specific outputs, which rely heavily on user input[42]. Thus, Generative AI applications might also warrant an individual and independent assessment similar to what the CJEU did in YouTube and Cyando for content storing and sharing applications[43].

Fourth, the policy argument[44] that Generative AI applications might not deserve liability immunities similar to those for traditional platforms due to their different nature merits attention. However, it may not be decisive in this context, as the issue of fake news, misinformation, disinformation, and content misuse is not exclusively about liability. Even critics of applying the DSA to Generative AI tools acknowledge the usefulness of the DSA's content moderation mechanisms against misleading, fake, or harmful AI-generated content. The DSA offers a potentially elegant solution, as Recital 41 suggests that the immunity regime and content moderation rules should be assessed independently[45]. Thus, irrespective of whether

perform a 'communication to the public' is related to their intermediary status under art. 6 of the DSA (the case was decided under art. 14 of the Electronic Commerce Directive, equivalent to today's art. 6 of the DSA). Therefore, even though content-sharing platforms index content, provide search tools, and suggest content based on user preferences, the Court affirmed they neither perform a 'communication to the public' nor lose their intermediary status. Thus, the argument of Hacker et al. (supra note 21) overlooks the nuances of CJEU intermediary case law.

41 As neatly put by BOTERO ARCILA, supra note 30, p. 484: "... ChatGPT's answers are assembled using algorithms that predict what makes sense as the next word, based on a user's prompt. OpenAI, thus, does not have any knowledge or an active role in controlling the content ChatGPT generates. It could thus be argued that its role is neutral in a similar way to how YouTube is neutral in hosting user-generated content …".

42 On the decisive nature of user prompts in the context of Generative AI applications see PATEL, David, "Artificial Intelligence & Generative AI for Beginners", Kindle Edition, 2024, p. 118-139.

43 For an overall assessment of Generative AI in the context of US law, specifically § 230 of 47 U.S.C., a provision not too dissimilar from the discussion in this paper under EU law, see BAMBAUER, Derek; SURDEANU, Mihai, "Authorbots", Journal of Free Speech Law, 2023 3(2), pp. 375-388. Their analysis demonstrates the nuanced reality of whether Generative AI applications can be considered content creators or mere tools serving the users who dictate the parameters of the output through their prompts.

44 As put forward by Edwards et. al, supra note 26.

45 Recital 41 of the DSA reads as follows: "In that regard, it is important that the due diligence obligations are adapted to the type, size and nature of the intermediary service concerned. This Regulation therefore sets out basic obligations applicable to all providers of intermediary services, as well as additional obligations for providers of hosting services and, more specifically, providers of online platforms and of very large online platforms and of very large online search engines. To the extent that providers of intermediary services fall within a number of different categories in view of the nature of their services and their size, they should comply with all the corresponding obligations of this Regulation in relation to those services. Those harmonised due diligence obligations, which should be reasonable and nonarbitrary, are needed to address the identified public policy concerns, such as safeguarding the legitimate interests of the recipients of the service, addressing illegal practices and protecting the fundamental rights enshrined in the Charter. **The due diligence obligations**

Generative AI applications are granted or denied immunity, the DSA's content moderation provisions could still apply[46]. In other words, the inclusion of Generative AI applications in the DSA does not guarantee immunity; if Generative AI tools behave as active service providers exceeding the threshold set by the intermediary phenotypes of the DSA, they will be denied immunity. Crucially, nonetheless, they will still have to comply with EU-wide content moderation obligations. Therefore, exploring how to integrate Generative AI tools under the DSA might be a reasonable alternative to their complete exclusion.

At the other end of the spectrum, more flexible approaches that consider applying the DSA to Generative AI applications carry some merits, especially in view of the fact that they represent a more nuanced approach to the problem of applying the DSA intermediary taxonomies to Generative AI applications. They may, nonetheless, need further development. One proponent acknowledges that classifying Generative AI tools as search engines is only a partial solution and may not adequately address content misuse[47]. The multimodal nature and extensive capabilities of Generative AI applications raise questions about whether they could fall under various intermediary service categories within the DSA. This requires further exploration and adjustment to avoid regulatory overload, uncertainty, and reduced effectiveness.

4. CONCLUSION

In conclusion, Generative AI harbours the potential to fabricate fake news, misinformation, and disinformation, thereby presenting substantial challenges for content moderation in the digital age. The applicability of the DSA to Generative AI applications has sparked a spectrum of opinions within the legal discourse. Some commentators maintain that the scope of the DSA is too restrictive to encompass these sophisticated tools, while others advocate for a more elastic interpretation, suggesting that Generative AI applications could be classified as search engines or hosting providers.

Despite these divergent viewpoints, there is a common consensus that under certain circumstances, Generative AI applications can indeed be regulated within the framework of the DSA. For instance, they might come under the purview of the DSA if integrated into traditional content-sharing platforms. Additionally, the DSA could be applied at a secondary level when AI-generated content is disseminated through these platforms, thereby ensuring a measure of regulatory oversight. This reflects a shared understanding that the existing legis-

are independent from the question of liability of providers of intermediary services which need therefore to be assessed separately *(emphasis added)*".

46 On the relationship between the immunity from liability and the due diligence obligations of digital service providers under the DSA see SCHWEMER, Sebastian: "Digital Services Act: a reform of the e-Commerce Directive and much more", in: SAVIN, Andrej; TRZASKOWSKI (eds.): Research Handbook on EU Internet Law, 2nd ed., Cheltenham: Edward Elgar, 2023, pp. 232-252, especially 244-245.

47 Botero Arcila supra note 30, p. 483-487.

lative framework possesses some flexibility to accommodate the unique challenges posed by AI-generated content.

However, the challenges involved in effectively regulating Generative AI necessitate further scholarly investigation. A wholesale dismissal of the possibility of incorporating Generative AI tools under specific DSA phenotypes, which entail content moderation and due diligence obligations, might not be the most effective approach. It is imperative to explore which intermediary phenotype within the DSA is most suited for the incorporation of Generative AI applications. Given the multimodal, complex, and hybrid capabilities of these applications, it might even be appropriate to consider a combination of different phenotypes.

Moreover, it is crucial to determine the appropriate mix of liability, potential immunity, and content moderation provisions that will ensure a safer digital environment. This endeavour requires a nuanced understanding of the dynamic interplay between regulation and innovation. The positions expressed in this paper highlight the necessity of further research and a more detailed examination of whether to integrate Generative AI into the DSA framework and if so how, without precipitating regulatory overload or engendering uncertainty. Such an exploration must necessarily aim at developing a coherent and effective regulatory strategy that addresses the unique challenges posed by Generative AI, while maintaining the delicate balance required to foster technological innovation and safeguard user protection in the digital realm.

Capítulo 7

DEEPFAKES Y DERECHO PENAL: DERECHO AL HONOR, A LA INTIMIDAD Y A LA PROPIA IMAGEN

María Isabel Montserrat Sánchez-Escribano[1]
Profesora Titular Laboral
Universidad de las Islas Baleares

RESUMEN: Durante los últimos meses, los medios de comunicación se han hecho eco de multitud de casos de manipulación de medios digitales a través de la técnica del Deepfake. El problema en sí no es la propia manipulación informática para crear contenido falso de una persona (durante los últimos años se ha empleado esta técnica de forma perfectamente legal) sino su utilización perniciosa para crear contenido sexual artificial –imágenes o vídeos pornográficos no reales ni consentidos– de mujeres y niñas, recrear a los ex en el metaverso, vestir o desvestir mujeres, o eliminar su individualidad para someterlos a los cánones tradicionales…

¿No es esto un atentando contra su libertad, intimidad y propia imagen? ¿y no es el acto de desvestir o vestir a una mujer una nueva forma de violencia sexual o machista? Sobre todo porque los estudios que se han realizado hasta el momento apuntan a que el 99% de los Deepfakes de contenido sexual afectan a mujeres y que entre los años 2022 y 2023 esta conducta se ha incrementado en un 429% (BIGAS FORMATJÉ, 2023: UOC.EDU). A realizar un análisis de estas cuestiones se dedica este texto.

PALABRAS CLAVE: Deepfake, Inteligencia Artificial, Derecho Penal, Honor, Intimidad, Propia Imagen

1 Este trabajo ha sido realizado en el marco del proyecto de investigación PID2022-140944OA-109, cuyo título es "Inteligencia artificial y Derecho: Análisis de la responsabilidad de los daños derivados del uso de sistemas de inteligencia artificial", financiado por el Ministerio de Ciencia e Innovación, la Agencia Española de Investigación y Fondos FEDER de la UE (MCIN/AEI/10.13039/501100011033/FEDER, UE).

1. INTRODUCCIÓN

La Inteligencia Artificial y, especialmente, la Inteligencia Artificial Generativa ha supuesto una revolución sin precedentes en la tecnología conocida hasta el momento (Natural Machine Intelligence, 2021). En este contexto, una de sus principales aplicaciones, los Deepfakes, han difuminado la línea entre lo real y lo artificial, ya que permiten generar contenido de vídeo, imagen y texto completamente nuevo y único, lo que ha abierto un mundo de posibilidades para la expresión artística, la innovación científica y la creatividad (Kaswan, 2023).

Uno de los aspectos más importantes de la Inteligencia Artificial Generativa es que, desde la aparición de Chat GPT en 2022, resulta accesible a toda la sociedad. No obstante, el amplio número de aplicaciones positivas que tiene esta tecnología están quedando relegados a un segundo plano debido a los numerosos casos de usos no éticos que se está haciendo de ella. Concretamente, la generación no consentida de fotografías con un contenido sexual explícito –imágenes o videos pornográficos– de mujeres adultas y, más detenidamente, de menores de edad.

A mi juicio, el problema no reside en la propia tecnología, sino en la utilización inadecuada y dañina que se ha hecho de ella. Durante los últimos años, esta técnica ha demostrado su potencial y se ha empleado con éxito en multitud de ocasiones, sin generar críticas ni complicaciones, para resucitar o rejuvenecer virtualmente a figuras icónicas o para reconstruir voces históricas[2].

A lo largo de este capítulo, se analizará el concepto de Deepfake, cómo se elaboran, qué consecuencias pueden tener los usos inadecuados de los mismos y se hará referencia a la posibilidad de sancionar desde la perspectiva del Derecho civil y del Derecho Penal la difusión de determinados Deepfakes.

2. ¿QUÉ ES UN DEEPFAKE?

El término Deepfake es una contracción de "*deep learning*" (aprendizaje profundo) y "fake" (falso). La razón de que se haya utilizado este término (deep) y no otro radica en la técnica utilizada para crearlos, denominada aprendizaje profundo o *deep learning*, una rama de la Inteligencia Artificial que se enmarca dentro del área conocida como aprendizaje automático (*machine learning*)[3]. Concretamente, los Deepfakes se crean utilizando dos técnicas de

2 Ejemplo de ello son el anuncio de Cruzcampo "Con mucho de acento" protagonizado por la emblemática Lola Flores, o la reconstrucción de la voz de John Fitzgerald Kennedy para recrear el discurso que hubiera pronunciado en el Dallas Trade Mart el día que fue asesinado. En el ámbito del entretenimiento, Star Wars ha sido un terreno especialmente fértil para el Deepfake. En la película "Rogue One: A Star Wars Story" se utilizó esta técnica para rejuvenecer a Carrie Fisher como la princesa Leia y para revivir al actor Peter Cushing en el papel del Gran Moff Wilhuff Tarkin. En "The Mandalorian", Luke Skywalker también fue rejuvenecido gracias al Deepfake.

3 El *deep learning* se caracteriza por el uso de redes neuronales artificiales interconectadas en múltiples capas, las cuales trabajan en conjunto para procesar datos y tomar decisiones (véase Le Cun, 2015). Cada capa de una red neuronal profunda recibe información de la capa anterior, la procesa y la envía a la siguiente capa. Las primeras

aprendizaje profundo (*deep learning*): Las redes generativas antagónicas[4] y autocodificadores variacionales[5]. Desarrollemos un poco más este concepto.

2.1. Concepto de Deepfake

El Deepfake es una técnica de Inteligencia Artificial que permite manipular medios digitales para crear imágenes, vídeos o audios falsos de una persona, animal, lugar o ambiente[6]. Actualmente, estos sistemas son capaces de generar contenido con tal grado de precisión que se hace muy difícil distinguir la falsificación del contenido auténtico (Yu et. al., 2019).

2.2. Clases de Deepfake

Por lo que se refiere al ser humano, que es lo que aquí nos interesa, es posible emular expresiones faciales, movimientos corporales e, incluso, la voz humana de una forma extremadamente realista. Con carácter general, pueden identificarse tres tipos de Deepfakes relacionados con las personas:

capas extraen características básicas de los datos, como por ejemplo bordes y colores en imágenes, mientras que las capas posteriores combinan estas características en patrones más complejos y abstractos.

4 Las redes generativas antagónicas (GAN) son la técnica central en la creación de Deepfakes (véase Karpathy et. Al., 2016). Consisten en dos redes neuronales enfrentadas: un generador y un discriminador. El generador produce contenido falso o sintético y el discriminador evalúa la autenticidad de este contenido. Durante el entrenamiento, estas dos redes neuronales compiten entre sí: el generador busca crear contenido que engañe al discriminador, mientras que el discriminador busca detectar cualquier inconsistencia en el contenido generado. Esta competencia iterativa conduce a una mejora continua en la calidad del contenido generado, ya que el generador aprende a crear imágenes o vídeos cada vez más realistas y, en consecuencia, más difíciles de distinguir de los originales.

5 Los modelos variacionales autoencoder (VAE) tienen potencial para ser utilizados en la creación de Deepfakes, pero no son la técnica principal que se emplea actualmente debido a que la calidad de las imágenes sintetizadas depende en gran medida de la calidad y cantidad de datos de entrenamiento, por lo que, si estos son limitados o de baja calidad, las imágenes generadas pueden verse borrosas y poco realistas.

6 Es necesario diferenciar este concepto del de transferencia de estilo (véase Canham et. al., 2023), una técnica que se utiliza para aplicar el estilo visual de una imagen de origen a otra imagen de destino. Esta técnica se basa en que el estilo de una imagen, como su paleta de colores, texturas y patrones, puede separarse de su contenido, como los objetos y las formas representadas en la imagen. Al separar ambas cosas, este modelo permite tomar el estilo visual de una imagen y aplicarlo a otra, creando así una imagen nueva pero que conserva el contenido y estilo artístico de la imagen de origen. La transferencia de estilo puede ser utilizada en el contexto de la creación de un Deepfake, ya que estos no solo imitan la apariencia física de una persona, sino también su estética personal para crear un estilo visual distintivo. No obstante, una mera modificación estética de una imagen, una mera transferencia de estilo, no puede considerarse un Deepfake en sí mismo.

2.2.1. Deepfake de Imagen

Las imágenes Deepfake representan la vertiente más común de esta tecnología, que tiene la capacidad de modificar fotografías existentes o generar nuevas imágenes que parecen realistas pero que son completamente fabricadas. En este caso, el *deep learning* aprende y replica detalles minuciosos de una imagen, tales como la textura de la piel, el reflejo de la luz en los ojos y otros rasgos faciales, y permite alterar las características faciales, cambiar el entorno o incluso crear rostros de personas que no existen en la realidad.

2.2.2. Deepfake de voz

La tecnología Deepfake también es capaz de generar grabaciones de voz que imitan a una persona específica con gran precisión. En este caso, los modelos de *deep learning* se entrenan con grabaciones de la voz de una persona, capturando sus patrones de habla, entonación y acento. Una vez que el modelo ha aprendido estas características, puede generar nuevos fragmentos de audio que suenan como si hubieran sido pronunciados por la persona original.

2.2.3. Deepfake de vídeo

Los Deepfakes de vídeo son una de las aplicaciones más impresionantes y controvertidas de esta tecnología. El proceso comienza con la recopilación de múltiples imágenes y videos de una persona, que se utilizan para entrenar el modelo. Una vez entrenado, este puede generar secuencias de vídeo donde la cara de la persona se superpone a la de otra persona o se altera su expresión facial y se realiza una sincronización labial para que coincidan con un audio diferente.

3. DERECHO AL HONOR, A LA INTIMIDAD Y A LA PROPIA IMAGEN Y DEEPFAKES

La utilización para fines perniciosos de la técnica del Deepfake ha hecho nacer en la sociedad la necesidad de reclamar responsabilidad respecto de determinadas conductas. Ello exige delimitar cuáles son los derechos principalmente afectados por estas y si el menoscabo de los mismos está cubierto por la normativa existente o si, por el contrario, requiere de la aprobación de nuevas normas jurídicas. Lo primero, lo analizaremos en el primer punto de este apartado, lo segundo, en los puntos segundo y tercero.

3.1. LA AFECTACIÓN DEL DERECHO AL HONOR Y A LA PROPIA IMAGEN EN LOS DEEPFAKES

Los derechos de los que principalmente se habla en lo que concierne a los Deepfake son el honor y la propia imagen, así como también la libertad de expresión e información y la libertad de creación artística. Se trata, como vemos, de derechos fundamentales de las personas. En este texto nos centraremos en el análisis de los dos primeros con la finalidad de analizar posibles vulneraciones que puedan generar responsabilidad en el ámbito civil y penal.

3.1.1. Derecho al honor

El derecho al honor se encuentra recogido en el artículo 18.1 de la Constitución Española. Su contenido esencial tiene dos vertientes: por un lado, la consideración que uno tiene sobre sí mismo y, por otro, la reputación o prestigio de una persona en su ámbito personal o profesional[7]. En este contexto, un Deepfake puede ser fácilmente utilizado para crear contenido falso y dañino sobre una persona, haciendo que parezca que ha hecho o dicho algo inapropiado, ilegal o inmoral. Esto puede afectar negativamente la percepción pública de la persona y su reputación personal y profesional.

Además, la rapidez con la que se genera un Deepfake y la facilidad con la que esta técnica permite simular contextos y situaciones completamente irreales/falsos, la hacen ideal para difamar y minar ese prestigio o reputación de cualquier persona. La difusión de información falsa o engañosa sobre una persona a través de la técnica del Deepfake o acciones como mostrar a esa misma persona realizando actos ilegales o inmorales en público -consumo de drogas, actos sexuales, comisión de delitos, etc.- pueden dañar gravemente la reputación de una persona.

3.1.2. Derecho a la propia imagen

El derecho a la propia imagen se encuentra también reconocido en el artículo 18.1 de la Constitución Española. Su contenido esencial se escinde en dos facetas: una positiva, que atribuye a todas las personas la facultad de difundir o publicar su propia imagen —definida esta como la exteriorización de la figura humana— libremente, y una negativa, consistente en la posibilidad de controlar e, incluso, negarse a la reproducción o divulgación de su imagen.

Así como es fácil constatar que la divulgación de Deepfakes con ánimo difamatorio constituye una clara vulneración del derecho al honor, lo cierto es que más controvertido resulta afirmar el aten-

7 Resulta controvertido en la doctrina el reconocimiento de un derecho al honor de las personas jurídicas (VIDAL MARTÍN, 2007) y, especialmente, de las empresas mercantiles (TENORIO SÁNCHEZ, 2023).

tado contra la propia imagen. ¿Por qué? Pues si por ejemplo tomamos por caso los distintos casos de pornografía no consentida que han salido en la prensa, lo único que es real en esas fotografías es la cara de las personas que aparecen en ellas, ya que el desnudo que las acompaña es completamente falso. Una imagen simulada generada por la Inteligencia Artificial y que no se corresponde con la apariencia real de las personas. ¿El derecho a la propia imagen cubre esa parte simulada?

3.2. LA CALIFICACIÓN DE DETERMINADOS DEEPFAKES COMO INTROMISIÓN ILEGÍTIMA DESDE LA PERSPECTIVA CIVIL

Visto brevemente cómo puede afectar la creación y difusión de Deepfakes al derecho al honor y a la propia imagen, veamos si esos usos pueden ser calificados, además, como una intromisión ilegítima en los mismos desde la perspectiva del Derecho civil.

3.2.1. Derecho al honor

Tanto difundir información falsa o engañosa sobre una persona a través de un Deepfake, como mostrar a una persona en una situación que daña su dignidad (lo que puede suponer también usar su imagen) constituyen, a los efectos del Derecho civil, una intromisión ilegítima en el derecho al honor de la persona. Este tipo de intromisiones son, por tanto, ser perseguibles judicialmente *ex* artículo 7.7 Ley Orgánica 1/1982, de 5 de mayo, de protección civil del derecho al honor, a la intimidad personal y familiar y a la propia imagen. Este artículo reza *la imputación de hechos o la manifestación de juicios de valor a través de acciones o expresiones que de cualquier modo lesionen la dignidad de otra persona, menoscabando su fama o atentando contra su propia estimación*. En este sentido, ni que decir tiene que el hecho de que se trate de un personaje con proyección pública no es óbice para tolerar la difusión de 'Deepfakes' que sitúen al protagonista en una escena que nunca tuvo lugar y que podría manchar su reputación".

3.2.2. Derecho a la propia imagen

El artículo 7.5 de la mencionada Ley Orgánica 1/1982, de 5 de mayo, establece que *[L]a captación, reproducción o publicación por fotografía, filme, o cualquier otro procedimiento, de la imagen de una persona en lugares o momentos de su vida privada o fuera de ellos, salvo los casos previstos en el artículo octavo, dos*. Este artículo hace referencia a varios supuestos en los que no se considera intromisión ilegítima:

a) *Su captación, reproducción o publicación por cualquier medio cuando se trate de personas que ejerzan un cargo público o una profesión de notoriedad o proyección pública y la imagen se capte durante un acto público o en lugares abiertos al público*.

b) La utilización de la caricatura de dichas personas, de acuerdo con el uso social.

c) La información gráfica sobre un suceso o acaecimiento público cuando la imagen de una persona determinada aparezca como meramente accesoria.

En el caso de los Deepfakes, cabe mencionar también el apartado 6 del primer artículo mencionado, que considera también intromisión ilegítima *[L]a utilización del nombre, de la voz o de la imagen de una persona para fines publicitarios, comerciales o de naturaleza análoga*.

En el ámbito del Derecho civil, la mera creación de un Deepfake usando la cara de una persona constituye una reproducción no consentida de la imagen y, por tanto, una intromisión ilegítima. Además, la utilización de la imagen de una persona obtenida con consentimiento para crear un determinado Deepfake pero con un fin diferente al que se han consentido también constituye una vulneración directa de este derecho y, por tanto, una intromisión ilegítima a los efectos de dicho precepto (véase Trujillo Cabrera, 2024). Ello, en mi opinión, con la excepción de que se trate de una caricatura de una persona pública.

La pregunta que surge ahora es: carecen de relevancia los casos en que se reproduce o publica la imagen de una persona pública en un acto público o en lugares abiertos al público, pero… ¿debe aplicarse la misma solución jurídica lo mismo si creo un Deepfake de Pedro Sánchez en un acto público de Corea del Norte? Recuérdese que una de las características del Deepfake es, precisamente, la dificultad que actualmente tenemos para determinar la veracidad de una imagen o un vídeo. En consecuencia, desde nuestro punto de vista, está claro que un Deepfake en el que se sitúa a un personaje público en un contexto o situación falsa con fines ilegítimos debe ser considerado una intromisión ilegítima en el derecho a la propia imagen.

4. LA (NECESARIA O NO) TIPIFICACIÓN DE DETERMINADAS CONDUCTAS DE DIFUSIÓN DE DEEPFAKES COMO DELITO

A la hora de valorar el encaje de las conductas de Deepfake en el Código Penal debe adelantarse que el 13 de octubre de 2023 se publicó en el Boletín Oficial de las Cortes Generales la *Proposición de Ley Orgánica de regulación de las simulaciones de imágenes y voces de personas generadas por medio de la inteligencia artificial*, presentada por el Grupo Parlamentario Plurinacional SUMAR, que tiene la finalidad de combatir los usos inadecuados de la técnica Deepfake. Esta propuesta pretende modificar un total de siete normas legales de nuestro ordenamiento jurídico y, entre estas modificaciones, una de las más importantes es la tipificación de las conductas de creación de Deepfakes con ánimo difamatorio como una nueva modalidad de delito de injurias, específicamente de injurias con publicidad[8].

8 Concretamente, su artículo 2 modifica el artículo 211 del Código Penal, añadiendo un párrafo segundo, quedado redactado el artículo de la siguiente forma:
«Artículo 211.
La calumnia y la injuria se reputarán hechas con publicidad cuando se propaguen por medio de la imprenta, la radiodifusión o por cualquier otro medio de eficacia semejante.
Salvo previa autorización expresa de la persona o personas afectadas, las simulaciones de imágenes, vídeos o audios de voz de estas generados a través de sistemas automatizados, software, algoritmos o mecanismos de inteligencia artificial que fueran difundidos a través de redes sociales serán consideradas como injurias hechas con publicidad.»

Sin embargo, la propuesta de Ley Orgánica no incluye ningún otro precepto que contemple posibles atentados constitutivos de delito de calumnias a través de Deepfakes o difusión de vídeos, imágenes o audios que utilicen la voz e imagen de la persona. Es decir, tampoco contempla ninguna reforma de los artículos 205 (delito de calumnias) ni del 197 del Código penal (delito de descubrimiento y revelación de secretos).

4.1. Derecho al honor

En relación con el derecho al honor, a la hora de analizar si la difusión de un Deepfake puede ser una conducta delictiva, se ha de valorar si la acción que realiza la persona que en él se muestra tiene o no relevancia penal, pues en este caso se estará imputando falsamente un delito y, por tanto, quien difunde sería autor de un delito de calumnias castigado en los artículos 205 a 207 del Código Penal (o, incluso, de un delito de acusación y denuncia falsa del artículo 457 si ese vídeo se entregara a la autoridad competente para la persecución de los delitos). Aun siendo interesante profundizar en esta cuestión, dado la escasa extensión que debe tener este texto, no se harán más comentarios al respecto y se pasará a analizar de forma más detenida el delito de injurias, el cual además constituye el objeto de la propuesta de reforma.

Como es sabido, el delito de injurias, tipificado en los artículos 208 a 210 del Código penal, tiene como bien jurídico protegido el honor de las personas. Este delito castiga la lesión de la dignidad de otra persona, a través de una acción o de una expresión, de forma que menoscabe su fama o atente contra su propia estimación. La diferencia principal entre el ilícito civil y el penal radica en la gravedad de la conducta. Así, únicamente son constitutivas de delito las injurias que, por su naturaleza, efectos y circunstancias sean tenidas en el concepto público por graves.

Como indica BARRIO, la valoración de la relevancia penal de la conducta se basa en el daño provocado al interés individual del receptor, para lo que debe tenerse en cuenta el contenido del mensaje, su contexto y el canal comunicativo utilizado (2022, p. 4), así como la difusión realizada (2022, p. 8). A este efecto, los criterios habituales que se usan como elementos de ponderación son: la relevancia pública de la cuestión objeto del debate (*ratione materiae* o *personae*), la propia expresión utilizada y su relación con la idea que transmitir; la veracidad de la afirmación; los hechos que sean susceptibles de prueba y no una opinión personal; las expresiones no injuriosas, desproporcionadas ni denigrantes que van más allá del mero aspecto crítico; y los usos sociales (BARRIO, 2022, 7, citando a la STS 201/2019, de 3 de abril, FJ 3º).

En el ámbito de las redes sociales, para considerar que una conducta es merecedora de sanción penal se tienen en cuenta dos aspectos: la propia ofensa del mensaje y la difusión del mensaje. Habitualmente, la difusión a través de Internet, especialmente de las redes sociales, de un mensaje difamatorio de una persona se considera un atentado grave contra la víctima. dado, por ejemplo, el amplio número de seguidores de la cuenta (BARRIO, pág. 8).

Lo anterior resulta plenamente aplicable a la creación de Deepfakes, donde, teniendo en cuenta la naturaleza informática e intangible de los mismos, la difusión a través de internet

es la vía por antonomasia para hacerlos llegar al mayor número de destinatarios posibles. Sin embargo, en mi opinión, en la valoración de la gravedad de la conducta debe partirse de la idea de que la información no es veraz, lo que cambia el juicio de ponderación que debe realizar el juzgador. En este caso, este debe centrarse únicamente en el carácter denigrante o desproporcionado de la situación mostrada en el Deepfake y en la intención del autor.

4.2. Derecho a la propia imagen

Respecto al derecho a la propia imagen, este se encuentra protegido penalmente a través del delito de descubrimiento y revelación de secretos, recogido en el artículo 197 del Código penal[9]. La pregunta en este caso es si la conducta de manipular una imagen no íntima con el objetivo de convertirla en una imagen íntima para dañar a una persona y posteriormente difundirla pueden subsumirse en este delito.

Esto resulta especialmente relevante, dado que uno de los usos habituales del Deepfake en la actualidad se enmarca dentro de lo que se conoce como pornografía de venganza y que se estima que el 99% del contenido pornográfico Deepfake está protagonizado por mujeres[10]. Además, entre 2022 y 2023, la cantidad de pornografía Deepfake creada aumentó un 464%, pasando de 3725 vídeos en 2022 a 21.019 en 2023 (UOC, 2023, online).

En este caso, a diferencia de lo que sucede con el honor, tal y como bien indica JAREÑO LEAL, este delito protege el derecho a la propia imagen sólo frente a determinados ataques, y, además, únicamente protege la imagen verídica de la persona, es decir, "la que representa la verdad" (quedando excluida la protección de la que es fruto de una manipulación) (JAREÑO LEAL, 2024, 6). En consecuencia, los Deepfakes sexuales no tienen actualmente cabida en este precepto del Código penal y tampoco son objeto de la reforma, por lo que nos encontraríamos ante una laguna jurídica. Interesante, en este caso, resulta valorar si la difusión de estos vídeos no conforma ya un atentado contra el honor, perseguible por la vía de la injuria.

5. CONCLUSIONES

I. Durante los últimos años hemos sido testigos de multitud de casos de creación difusión no consentida de contenido discriminatorio, ofensivo, falso o engañoso a través de la utilización de la técnica del Deepfake. Aunque este hecho ha generado una importante desconfianza en esta tecnología, el problema no reside en la tecnología en sí misma, sino en la utilización inadecuada y dañina para los derechos de las personas que se ha hecho de ella.

9 Sobre este delito y su bien jurídico véase mi artículo (2015, 323-363).

10 El término pornografía de venganza o pornovenganza, proveniente del término en inglés pornrevenge hace referencia a la publicación no consentida de contenido sexual (vídeo o imagen) de una ex-pareja a través de Internet bien por despecho bien como vindicación por haber roto la relación. (Fiorio y Vetis, 2020)

II. Muchos de estos usos no éticos de los Deepfakes constituirán claramente un ilícito civil que afectarán a los derechos al honor y a la propia imagen de la persona afectada.

III. Sin embargo, en el ámbito del Derecho Penal existen muchísimos más problemas y ciertas lagunas para sancionar como delito las difusiones no consentidas de Deepfakes, ya que se trata de contenido no real de la persona.

6. BIBLIOGRAFÍA

ASHISH JAIMAN, Observer Research Foundation, Debating the ethics of Deepfakes, en Digital Frontiers, 2020, accesible en: https://www.orfonline.org/expert-speak/debating-the-ethics-of-Deepfakes.

BARRIO, Rodrigo Miguel, El delito de injurias y las redes sociales. El número de 'followers' y otras variables ambientales como elementos de valoración del daño, Revista de Estudios de Derecho y Ciencia Política, nº 36, 2022.

Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Fomentar un planteamiento europeo en materia de inteligencia artificial. COM/2021/205 final. Bruselas, 21 de abril de 2021.

D. CANHAM, T., MARTÍN, A., BERTALMÍO, M., PORTILLA, J., Using Decoupled Features for Photo-realistic Style Transfer, en Electrical Engineering and Systems Science, Cornell University, 2023, accesible en: https://arxiv.org/pdf/2212.02953 y https://www.io.csic.es/ciencias-de-la-imagen/transferencia-del-estilo-visual-de-una-fotografia-a-otra-imagen-o-video/

CHANDLER, D. Y MUNDAY, R., A Dictionary of Media and Communication, Oxford University Press, 2020, accesible en: https://www.oxfordreference.com/display/10.1093/acref/9780198841838.001.0001/acref-9780198841838?btog=chap&hide=-true&jumpTo=Narrow+AI&page=107&pageSize=20&skipEditions=true&sort=titlesort&source=%2F10.1093%2Facref%2F9780198841838.001.0001%2Facref-9780198841838 Doi:10.1093/acref/9780198841838.001.0001

Editorial, Much to discuss in AI ethics. Natural Machine Intelligence 4, 1055–1056, 2022. https://doi.org/10.1038/s42256-022-00598-x

Editorial, The big question. Natural Machine Intelligence 3, 737, 2021. https://doi.org/10.1038/s42256-021-00395-y

GAUDAMUZ, A., La inteligencia artificial y el derecho de autor, en OMPI Revista, 2017, accesible en: https://www.wipo.int/wipo_magazine/es/2017/05/article_0003.html

JAREÑO LEAL, Ángeles, El derecho a la imagen íntima y el Código penal La calificación de los casos de elaboración y difusión del deepfake sexual, Revista Electrónica de Ciencia Penal y Criminología, 26-09, 2024.

KARPATHY, A., ABBEEL, P., BROCKMAN, G., CHEN, P., CHEUNG, V., DUAN, R., GOODFELLOW, I., KINGMA, D., HO, J., HOUTHOOFT, R., SALIMANS, T., SCHUL-

MAN, J., SUTSKEVER, I. Y WOJCIECH, Z., Generative Models, OpenAI, 2016, accesible en: https://openai.com/index/generative-models/

KASWAN,K. S., DHATTERWAL, J. S., MALIK K. Y BALIYAN, A., Generative AI: A Review on Models and Applications, en 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), Greater Noida, India, 699-704, 2023. DOI: https://10.1109/ICCSAI59793.2023.10421601

LECUN, Y., BENGIO, Y. Y HINTON, G., Deep learning, en Nature 521, 436–444, 2015. DOI: https://doi.org/10.1038/nature14539

MARTÍN RÍOS, P., Empleo de big data y de inteligencia artificial en el ciberpatrullaje: de la tiranía del algoritmo y otras zonas oscuras, Revista de Internet, Derecho y Política, 36, 1-13, 2022.

MONTSERRAT SÁNCHEZ-ESCRIBANO, María Isabel, Libertad informática y protección de datos: desarrollo en la jurisprudencia del Tribunal Constitucional y tutela penal en el delito de descubrimiento y revelación de secretos, Anuario iberoamericano de justicia constitucional, ISSN 1138-4824, Nº. 19, 2015, págs. 323-363.

REUTERS, Explainer: What is Generative AI, the technology behind OpenAI's ChatGP, 2023, accesible en: https://www.reuters.com/technology/what-is-generative-ai-technology-behind-openais-chatgpt-2023-03-17/

RICHARD GONZÁLEZ, M., Los sistemas biométricos de reconocimiento facial en la Unión Europea en el marco del desarrollo de la Inteligencia Artificial, en Justicia, nº 1, 147-281, 2023.

SANCHO CAPARRINI, F., Variational AutoEncoder, en Blog Personal, 2024, accesible en: https://www.cs.us.es/~fsancho/Blog/posts/VAE.md

SIMÓ SOLER, E., Retos jurídicos derivados de la Inteligencia Artificial Generativa: Deepfakes y violencia contra las mujeres como supuesto de hecho, en InDret, 2, 2023. DOI: 10.31009/InDret.2023.i2.11

Stanford University - Human Centered Artificial Intelligence, Artificial Intelligence Index Report, 2024, accesible en: https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf

SWORNA, Z.T., URZEDO, D., HOSKINS, A.J. et al. The ethical implications of Chatbot developments for conservation expertise, en AI Ethics, 2024. https://doi.org/10.1007/s43681-024-00460-3

TENORIO SÁNCHEZ, P., Derecho al honor o reputación de las empresas mercantiles en la jurisprudencia del Tribunal Constitucional Español, del Tribunal Europeo de Derechos Humanos y del Tribunal de Justicia de la Unión Europea, en Anuario Iberoamericano de Justicia Constitucional, 27(2), 447-477, 2023. Doi: https://doi.org/10.18042/cepc/aijc.27.14

TRUJILLO CABRERA, C., El derecho a la propia imagen (y a la voz) frente a la inteligencia artificial, en InDret, 1, 74-113, 2024.

UNICRI y EC3, Malicious uses and abuse of Artificial Intelligence, Trend Micro Research, 2020, accesible en: https://unicri.it/sites/default/files/2020-11/Abuse_ai.pdf

UNESCO, Recomendación sobre la ética de la Inteligencia Artificial, 2021, accesible en: https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

URÍAS, J., El derecho a la producción y creación literaria, artística, científica y técnica, en Comentarios a la Constitución española: XXX aniversario, Ed. 1. Fundación Wolters Kluwer. 2008.

UOC, 'Deepfakes' pornográficos: Cuando la IA desnuda tu intimidad y vulnera tus derechos, 2023, accesible en: https://www.uoc.edu/es/news/2023/265-deepfakes-pornograficos-cuando-IA-desnuda-tu-intimidad-vulnera-tus-derechos

VIDAL MARTÍN, T., Derecho al honor, personas jurídicas y Tribunal constitucional, en La reforma del Tribunal Constitucional, 563-574, Valencia, Tirant lo Blanch, 2007.

YU, N., DAVIS, L. Y FRITZ, M., Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints, en 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 7555—7565, 2019. DOI: https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00765

Capítulo 8

APROXIMACIÓN AL USO DE LA INTELIGENCIA ARTIFICIAL EN LA PERFILACIÓN CRIMINAL

Nancy Carina Vernengo Pellejero
Profesora Lectora de Derecho Procesal
Universitat de Barcelona

RESUMEN: A principios de la década de los 80, los agentes de la *Behavioral Science Unit* del FBI plantearon una forma pionera de crear perfiles criminales: el uso de un programa informático de análisis de datos. Más adelante identificaríamos estas nuevas tecnologías con la Inteligencia Artificial aplicada a la investigación de delitos. Si bien aquellos pioneros de la perfilación criminal nos sugerían un embrionario sistema de análisis de datos como fórmula para agilizar la investigación criminal, actualmente el uso de estas nuevas tecnologías (también, y muy especialmente en la Administración de justicia) nos plantean diversas dudas sobre su verdadera utilidad y sus límites. Este artículo representa una aproximación a los avances en el uso de la IA sobre la investigación de delitos, especialmente en lo relativo a la perfilación criminal y a la actividad predictiva; así como los pros y contras que nos sugiere su utilización, atendiendo a su regulación actual. Del análisis de estas cuestiones podremos concluir que, si bien el uso de las nuevas tecnologías puede resultar de gran ayuda en algunos casos, también presenta en su haber ciertos límites que nos hacen poner en duda su eficiencia real.

PALABRAS CLAVE: Inteligencia Artificial – Perfiles criminales – Criminalística – Algoritmos – Proceso penal

1. INTRODUCCIÓN

La injerencia de la Inteligencia Artificial no está (ni seguramente estará) exenta de debate, independientemente del uso y la finalidad que se le asocien. Hablar de la Inteligencia Artificial nos lleva a considerar el alcance de sus diversos ámbitos de aplicación, los cuales están directamente relacionados con la amplia gama de recursos tecnológicos que engloba (aprendizaje computacional; reconocimiento de voz; procesamiento de datos,…). Partiendo de esta base, las aplicaciones en IA se nos muestran casi infinitas en su haber, sin perjuicio de los límites que se nos plantean en torno a cuestiones tan delicadas como el tratamiento de los datos personales, o la debida observancia de los derechos fundamentales de las personas. Trasladadas al ámbito propio de la Administración de Justicia, y los diversos usos a que se está destinando esa Inteligencia Artificial; estas cuestiones problemáticas se tornan aún más complejas, si tenemos presente que su uso se está implementando, incluso, en la investigación criminal. En los próximos apartados trataremos de ahondar en la imparable presencia de la IA en el ámbito judicial, prestando especial atención a su incorporación como método para la investigación de delitos violentos, y especialmente, como herramienta predictiva en la construcción de perfiles criminales.

2. INTERFERENCIA DE LA INTELIGENCIA ARTIFICIAL EN LA ADMINISTRACIÓN DE JUSTICIA

Agigantados han sido los pasos de la IA desde que sus creadores (entre ellos, John McCarthy, Marvin Minsky y Seymour Papert) revelaran sus avances sobre esta materia al mundo. Muchos son los usos que actualmente damos a la IA, y la Administración de Justicia, en sus distintos ámbitos, también ha sido testigo de ello. No podemos obviar las ventajas que puede representar su uso en entornos tan delicados como la detección de situaciones de riesgo sobre las mujeres víctimas de violencia de género (Viogén); o las aplicaciones puestas a disposición de los juristas por empresas del sector jurídico, basadas en justicia predictiva, y que permiten al abogado prever el posible resultado que obtendrá de las actuaciones promovidas ante un Tribunal; así como el sistema Veripol, para la detección de denuncias falsas (art. 457 CP) [1]. Pero, por otro lado, y a pesar de su aparente utilidad, algunos de estos programas han acabado cayendo en desuso, o han sido progresivamente retirados por las administraciones que los utilizaban, debido a defectos en su configuración, o por simple incompatibilidad con los derechos e intereses de los ciudadanos. Esta realidad derivaría, en cierto modo, de las dudas generadas respecto a la recopilación y tratamiento de los datos personales por parte de las empresas privadas encargadas de crear y administrar estos sistemas. Es el caso, entre otros especialmente destacables, de EuroCop PredCrime, un sistema basado en el tratamien-

1 MONTESINOS GARCÍA, Ana: "Algoritmos predictivos y perspectiva de género en el proceso penal", Revista d'Internet, Dret i Política (2023) (39), pp. 1-12.

to masivo de datos asociados sobre zonas de alta criminalidad (o mapas de calor), a partir del cual las FCSE podrían predecir los lugares en los cuales se podría requerir más su presencia (lo que permitiría, en teoría, administrar mejor sus medios). Huelga decir que la información manejada por el sistema en cuestión provenía, mayoritariamente, de las bases de datos policiales. Ya sea por las dudas que se suscitaron respecto al tratamiento de esta información confidencial, o por la verdadera utilidad de este tipo de sistemas (algunos de ellos tachados directamente de discriminatorios, por el hecho de marcar como zonas conflictivas los barrios donde se localiza un mayor número de personas de una etnia o nacionalidad específica); lo cierto es que algunos de estos sistemas no han conseguido su plena implementación en el sistema policial, ni judicial, a pesar de la inversión que cada una de estas operaciones representó para las arcas del Estado. Sin perjuicio de las dudas que algunos de estos sistemas pueden haber llegado a generar, no podemos negar la utilidad de otros (como VioGén y otros sistemas basados en el uso de *bots*), y su necesaria implementación para automatizar procesos, a veces meramente burocráticos, en situaciones que requieren de un tratamiento agilizado.

2.1. Investigación criminal y nuevas tecnologías

A lo largo de los últimos cien años, la investigación criminal ha sido testigo de la incorporación de novedosas técnicas que han representado un auténtico avance en la criminalística. Actualmente las ciencias forenses representan el paradigma del avance de las nuevas tecnologías aplicadas a la investigación de delitos; máxime en lo que a cibercrimen se refiere. Entre las diversas aplicaciones que se han dado a estas herramientas tecnológicas, destacan especialmente las relacionadas con la policía predictiva, amén de su relación con la perfilación criminal y el análisis de datos asociados a la identificación de presuntos delincuentes. Sin perjuicio de las ventajas que pueden derivarse de un sistema aparentemente automatizado, la realidad nos ha mostrado cuan compleja puede resultar su implementación, si tenemos presente la posible vulneración de los derechos fundamentales de los ciudadanos.

2.2. IA, justicia predictiva y perfilación criminal

Entre las iniciativas legislativas promovidas en la última década, destacan las dirigidas a regular la aplicación de las nuevas tecnologías en la Administración de Justicia; y, especialmente, conviene remarcar aquellas dirigidas a la aplicación de la IA como herramienta predictiva. En todo caso, y con independencia de la aparente utilidad que las fórmulas predictivas puedan representar en nuestra Administración de Justicia, lo cierto es que aún no disponemos de un tratamiento uniforme sobre esta cuestión, ni conocemos el verdadero alcance de su influencia en el enjuiciamiento criminal[2]. Tampoco en lo que respecta a la denominada

2 BORGES BLÁZQUEZ, Raquel: *Inteligencia Artificial y proceso penal*, Navarra: Aranzadi, 2021.

"policía predictiva", y la posibilidad de avanzarse a la comisión de un delito utilizando estas nuevas tecnologías. Una problemática que se extiende más allá de las fronteras de nuestro país, ahondando aún más si cabe en la necesidad de una regulación específica sobre el uso y los límites de estos sistemas

2.2.1. La importancia de la perfilación en la investigación criminal

La perfilación criminal parece estar en auge desde las últimas décadas. Numerosas y constantes son las alusiones a esta técnica de investigación en los medios de comunicación, el cine, o la literatura, entre otros. Sin embargo, conviene recordar que, a pesar de su copiosa mención en distintos medios, no se trata de una técnica novedosa, ni su implementación en la investigación criminal es reciente. Ya en el siglo XIX, Thomas Bond y George Phillips, incorporaron el arte de la perfilación criminal en la investigación del caso de Jack el Destripador (no resuelto aún); para lo cual tomaron como punto de partida las conclusiones arrojadas por los forenses en las autopsias de las víctimas (lógicamente, no comparables a las técnicas actualmente aplicadas en la medicina forense), junto a evidencias obtenidas en las propias escenas de los crímenes; y, por descontado, algo de intuición y deducción[3]. Sin perjuicio de la utilización de ciertos métodos asociados a la perfilación criminal en algunos casos específicos a lo largo de las décadas siguientes, el salto cualitativo vino de la mano de la Unidad de Análisis de Conducta del FBI; cuyos agentes no solamente reordenaron y reformularon las técnicas utilizadas hasta el momento (que originariamente carecían de unas pautas o protocolos específicos), sino que también ayudaron a incorporar nuevos conocimientos sobre la materia, que aún hoy son utilizados por agentes de policía de todo el mundo[4].

De este modo, a principios de la década de los 70 los agentes especiales del FBI comenzaron a aplicar los principios de las ciencias psicológicas al análisis de la conducta humana. Ello los llevó a establecer ciertas pautas en el comportamiento de los delincuentes que, entre otras cosas, les permitió plantear una primera clasificación entre delincuentes organizados y desorganizados, o mixtos[5]. A partir de aquí, el desarrollo de la técnica de perfilación criminal

3 GONZÁLEZ ARRIETA, Angélica; GIL GONZÁLEZ, Ana; DE LUIS REBOREDO, Ana: "Inteligencia Artificial en construcción de perfiles de asesinos en serie", en: Vicente Gabriel, Jorge; Arberas, Iñaki (coord..): *Avances en informática y automática. Decimocuarto Workshop*, Salamanca: Universidad de Salamanca, 2020, pp. 2-27.

4 Una de las técnicas pioneras en este campo consistió precisamente en las entrevistas privadas de los agentes con distintos criminales (algunos de ellos asesinos en serie), con el objetivo de obtener información sobre su comportamiento, los posibles orígenes de su actividad criminal, y su motivación. Aunque comenzaron como una relación de entrevistas informales, en las cuales no se seguían pautas preestablecidas o recogidas en ningún protocolo específico, los agentes fueron capaces, tras varios años de entrevistas, de crear un manual de clasificación de los delincuentes, que aún hoy sigue siendo un referente en la investigación de crímenes violentos. Para una visión más amplia de este particular: DOUGLAS, John; BURGESS, Ann; RESSLER, Robert: *Crime classification manual*, New York: Lexington Books, 1992.

5 RESSLER, Robert; BURGESS, Ann; DOUGLAS, John: *Sexual homicide. Patterns and motives*, New York: Lexington Books, 1995, pp. 9-11. Especialmente ilustrativas sobre esta cuestión (en particular, sobre la perfilación de delin-

ha sido imparable, aplicándose a todo tipo de delitos de forma aparentemente efectiva (delitos sexuales; homicidios; violencia de género; etc.). No obstante, debemos tener presente que buena parte de esta tarea policial, más allá del análisis de los datos recogidos en las escenas de los crímenes, requiere también de una combinación entre experiencia e intuición; unidas, asimismo, a la aplicación de otras evidencias policiales y forenses, imprescindibles en toda investigación criminal. Probablemente una parte de esta problemática podría quedar resuelta con el uso de la IA y la minería de datos aplicada a la creación de perfiles criminales, lo que automatizaría un proceso que requiere de una constante inversión de tiempo y medios humanos, no siempre disponibles. Sin embargo, debemos tener presente que el hecho de automatizar un procedimiento como la perfilación criminal, solamente representa una parte de la solución. Queda abierta, irremediablemente, la cuestión relativa a la debida observancia de los derechos fundamentales de las personas; una cuestión aún no resuelta en nuestra legislación (más allá de lo recogido en la Ley de Protección de datos, o en la propia CE), y que la normativa comunitaria pretende regular con mayor o menor acierto; como fue en su momento la Directiva (UE) 2016/681, del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativa a la utilización de datos del Registro de nombres de los pasajeros, para la detección, prevención, investigación y enjuiciamiento de delitos graves y de terrorismo), a fin y efecto de identificar aquellos pasajeros en tránsito con un alto índice de riesgo para la seguridad aérea; y de forma más general, en el Reglamento Europeo de IA, cuyo Título III, Capítulo I, se dedica a los sistemas de IA de Alto Riesgo (HRAIS), entre los que se incluyen los utilizados por los CFSE en su actividad predictiva y de investigación.

2.2.2. La irrupción de las nuevas tecnologías en la actividad predictiva: la IA al servicio de la perfilación criminal

Superado cualquier atisbo de ficción novelística, la realidad nos ha golpeado con una auténtica cascada de nuevas técnicas en la investigación criminal, entre las que destaca, especialmente, el uso de sistemas de Inteligencia Artificial en la construcción de perfiles criminales. Es preciso tener en cuenta, en todo caso, que la injerencia de esta IA en el ámbito de la investigación criminal se remonta a décadas atrás; teniendo como antecedente más destacable uno de los primeros casos documentados hacia el año 1983. En esa ocasión, los agentes de la Unidad de Análisis de Conducta del FBI consiguieron construir el perfil criminal de un pirómano que actuaba en lugares de culto de Nueva Inglaterra, a partir de las evidencias halladas en el lugar de los hechos, combinados con otros datos asociados a su experiencia en casos similares anteriores. El resultado de esta suerte de técnica pionera fue una descripción altamente detallada del presunto delincuente, en la que se incluía, además, el posible lugar

cuentes sexuales), de nos muestran los trabajos de HAZELWOOD, Robert: "Analyzing the rape and profiling the offender", en: Hazelwood, Robert; Burgess, Ann Wolbert (ed.): *Practical aspects of rape investigation*, Segunda edición, Boca Ratón-Nueva York: CRC Press, 1995, pp. 155-181.

donde localizarlo, posibilitando así su detención y posterior confesión de los hechos[6]; todo un hito en aquel momento. Este primer antecedente del uso de la IA para la creación de perfiles criminales derivó en su momento en la creación de un programa informático de perfilación, con base en una recopilación exhaustiva de patrones de comportamiento criminal. A partir de aquí, surgió el denominado Arson Information Management System (AIMS), utilizado por el Centro Nacional para el Análisis de Crímenes Violentos del FBI (NCAVC, en sus siglas en inglés); que permitió predecir, de forma concreta, las fechas y localizaciones en las que podrían cometerse futuros actos delictivos. En definitiva, el germen de lo que hoy conocemos como "policía predictiva", orientada tanto a la investigación de delitos violentos, como a su prevención.

Considerando estos antecedentes, lo primero que debemos tener presente es que, con independencia de la aparente falta de medios con los que se contaba en ese primer caso documentado, actualmente la IA ha conseguido abrirse paso entre las distintas herramientas y técnicas al servicio de la investigación; con un importante peso en la actividad predictiva. Es menester matizar, en cualquier caso, que se trata de un ámbito en constante evolución dentro de las técnicas de investigación criminal y, en consecuencia, carecemos de una relación cerrada de recursos a disposición de las FCSE.

Técnicamente, los sistemas utilizados en la investigación y la prevención de delitos violentos se basan, por una parte, en el "aprendizaje automático", en combinación con el denominado "aprendizaje profundo"; y, por extensión, con el "aprendizaje de refuerzo". Partiendo de estos tres métodos de aprendizaje, se ha configurado toda una red de sistemas de información, que bebe directamente de los datos proporcionados por organismos jurídicos y policiales, y cuyo objetivo primordial pasa por la detección de actividad criminal en determinados entornos, para promover la adopción automática de decisiones (considerando el auge de la cibercriminalidad en esta última década, en la actualidad resulta imprescindible contar con estas técnicas de investigación). Para ello, y además de la integración de datos (la mayoría de ellos obtenidos mediante bases de datos policiales, pero sin olvidar también la información obtenida por otros medios; como las cámaras con identificación facial; los registros de huellas dactilares o de ADN)[7], la perfilación criminal basada en IA, utiliza tanto el método inductivo, como deductivo, en la construcción de estos perfiles. Partiendo del método inductivo, se recoge y analiza la información de las bases de datos generales de delincuentes, para alcanzar las correspondientes conclusiones con base en el método deductivo (o pensamiento descendente)[8]. De este proceso, con base en estos sistemas algorítmicos, surgirá una clasificación de los ciudadanos, de acuerdo con la conducta que manifiesten y su potencial peligrosidad; lo que resulta, cuanto menos, discriminatorio.

6 ICOVE, David: "Automated crime profiling", FBI Law Enforcement Bulletin, (1986) 12 (55), pp. 27-30.

7 CUATRECASAS MONFORTE, Carlota: *La Inteligencia Artificial como herramienta de investigación criminal*, Madrid: La Ley, 2022, pp. 147-149.

8 GOYAL, Aditya; GUPTA, Aime; SHAH, Alisha; ALEXANDER, Meyga Anne; AARTHI, N. "Criminal profiling using machine learning", Internationa Research Journal os Engineering and Technology (2020) 7 (6), p. 6332.

En la práctica, esta tarea asociada a la perfilación y la prevención de delitos se ha traducido en distintos programas; como el "Gang Violence Matrix", aplicado por la Policía Metropolitana de Londres en 2012, para la detección y prevención de actividades delictivas por parte de las pandillas que merodeaban en determinadas zonas de la ciudad, partiendo, claro está, del perfil concreto de los miembros de estas bandas (origen, edad, ocupación, nivel de estudios, etc.)[9]. En Alemania, la lucha contra la delincuencia organizada y las bandas terroristas, les llevó a implementar un sistema como RADAR-iTE 2.0, para la detección de terroristas en potencia. Y en similares términos, pero a nivel del espacio Schengen, encontramos el "Visa Information System (VIS)"; basado en el intercambio de datos sobre las personas que transitan por los Estados del espacio Schengen, tanto si provienen de la UE, como de terceros países; con el objetivo de neutralizar a posibles delincuentes y prevenir la actividad criminal. Estos sistemas, más allá de la efectividad que pudieran presentar, no nos alejan de la polémica que se deriva de la implementación de una herramienta dirigida a un colectivo o lugar en concreto (un factor altamente discriminatorio); del mismo modo que se vulneran otros derechos (como la presunción de inocencia). La propia Agencia de Derechos Fundamentales de la UE (FRA) se ha pronunciado al respecto, remarcando la necesidad de posponer el uso de estas técnicas de elaboración de perfiles criminales, y otros programas orientados a la prevención, hasta que contemos con una regulación específica y garantista de los derechos de los ciudadanos, amén de la necesaria evaluación periódica de la definición legal de los términos asociados al uso de la IA[10].

3. CONCLUSIONES

Si bien cualquier cuestión que pueda representar un avance para la sociedad, sobre todo en lo relativo a la seguridad y la justicia, debería representar un factor del todo positivo, en el caso de la implementación de la IA cabe plantear, sin duda, ciertas reservas. Uno de los principales dilemas que presenta el uso de la IA debemos asociarlo, irremediablemente, a la posible afectación que ostenta respecto a la afectación de la esfera privada de los ciudadanos y sus derechos fundamentales. Cualquiera que sea la finalidad de los sistemas configurados a partir de la IA, y puestos a disposición de las Administraciones y la ciudadanía, nos obliga a establecer ciertos límites en su aplicación, en aras de salvaguardar los derechos de los ciudadanos. Sin embargo, no resulta sencillo poner límites, cuando las aplicaciones de la IA son tan imprevisibles como raudas en su constante evolución. La normativa va un paso por detrás de su efectiva implementación en distintos ámbitos, lo que en la práctica se traduce en un manifiesto problema de inseguridad jurídica.

9 A este programa le siguieron otros, como el "Qlik Sense", orientado también a la prevención de delitos.

10 EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS: *Construir correctamente el futuro. La Inteligencia Artificial y los derechos fundamentales*, Luxemburgo: Oficina de Publicaciones de la Unión Europea, 2021, pp. 5-12.

Desconocemos a ciencia cierta el alcance que la IA llegará a tener en nuestras vidas en los próximos años, pero observando su rápida e imprevisible expansión en todos los ámbitos, es inevitable pensar en una presencia aún mayor en entornos que hasta ahora le habían sido vetados. Es preciso tener en cuenta, no obstante, que el funcionamiento de estos sistemas de IA tiene como base la minería de datos. El hecho que empresas privadas tengan en sus manos el acceso directo a información sensible de los ciudadanos, (con independencia de la causa que motive ese acceso y de las aparentes bondades que estos programas pueden acabar suponiendo para la seguridad y el bienestar de los ciudadanos), no nos aleja de una problemática tangible y alarmante: nuestra información personal está en manos de un grupo de empresas, cuya forma de tratarlos nos es absolutamente desconocida. Visto desde este punto de vista, resulta imprescindible, a todas luces, la debida observancia de los derechos fundamentales de los ciudadanos (intimidad, igualdad, presunción de inocencia,…); en el bien entendido que no podemos sacrificarlos en pro de una mayor efectividad en las tareas policiales. Ello nos obliga, en definitiva, a una mayor previsión legislativa sobre el uso de la IA en los distintos ámbitos de nuestro ordenamiento jurídico; y no solamente a nivel de la UE.

Las reflexiones anteriores nos arrastran a otra cuestión aún más alarmante si cabe, y es el hecho de que aquellas funciones estrictamente policiales y judiciales se están derivando a entidades privadas, con fines manifiestamente distintos a los propios de los CFSE y los operadores jurídicos. Llegados a este punto no podemos dejar de preguntarnos si realmente podemos concebir un sistema judicial y policial, en el cual las funciones que propiamente han de desarrollar profesionales de la Administración de Justicia, se dejan en manos de terceras personas (particulares), ajenas a los códigos profesionales y deontológicos propios de cualquier profesión jurídica, y que actúan movidos por intereses puramente comerciales y ajenos a la tarea policial.

4. BIBLIOGRAFÍA

BORGES BLÁZQUEZ, Raquel: *Inteligencia Artificial y proceso penal*, Navarra: Aranzadi, 2021.

CUATRECASAS MONFORTE, Carlota: *La Inteligencia Artificial como herramienta de investigación criminal*, Madrid: La Ley, 2022.

DOUGLAS, John; BURGESS, Ann; RESSLER, Robert: *Crime classification manual*, New York: Lexington Books, 1992.

EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS: *Construir correctamente el futuro. La Inteligencia Artificial y los derechos fundamentales*, Luxemburgo: Oficina de Publicaciones de la Unión Europea, 2021.

GIMENO BEVIÁ, Jordi: "Instrumentos actuales de policía y justicia predictiva en el proceso penal español: análisis crítico y reflexiones de lege ferenda ante aplicaciones futuras", Estudios penales y criminológicos (2023) (44), pp. 1-20.

GONZÁLEZ ÁLVAREZ, José Luis; SANTOS HERMOSO, Jorge; CAMACHO COLLADOS, Miguel: "Policía predictiva en España. Aplicación y retos futuros", Behaviour & Law Journal (2020) 6 (1), pp. 26-41.

GONZÁLEZ ARRIETA, Angélica; GIL GONZÁLEZ, Ana; DE LUIS REBOREDO, Ana: "Inteligencia Artificial en construcción de perfiles de asesinos en serie", en: Vicente Gabriel, Jorge; Arberas, Iñaki (coord..): *Avances en informática y automática. Decimocuarto Workshop*, Salamanca: Universidad de Salamanca, 2020, pp. 2-27.

GOYAL, Aditya; GUPTA, Aime; SHAH, Alisha; ALEXANDER, Meyga Anne; AARTHI, N. "Criminal profiling using machine learning", Internationa Research Journal os Engineering and Technology (2020) 7 (6), pp. 6331-6337.

HAZELWOOD, Robert: "Analyzing the rape and profiling the offender", en: Hazelwood, Robert; Burgess, Ann Wolbert (ed.): *Practical aspects of rape investigation*, Segunda edición, Boca Ratón-Nueva York: CRC Press, 1995, pp. 155-181.

ICOVE, David: "Automated crime profiling", FBI Law Enforcement Bulletin, (1986) 12 (55), pp. 27-30.

MONTESINOS GARCÍA, Ana: "Algoritmos predictivos y perspectiva de género en el proceso penal", *Revista d'Internet, Dret i Política* (2023) (39), pp. 1-12.

RESSLER, Robert; BURGESS, Ann; DOUGLAS, John: *Sexual homicide. Patterns and motives*, New York: Lexington Books, 1995.

Capítulo 9

LA UTILIZACIÓN DE LA INTELIGENCIA ARTIFICIAL PARA LA PRODUCCIÓN DE PORNOGRAFÍA: SU ENCAJE EN EL DERECHO PENAL ESPAÑOL.

Emilio Muñoz Campaña
emiliomucamp@correo.ugr.es
Universidad de Granada
Áreas temáticas: Derecho y criminología.

1. INTRODUCCIÓN.

La noticia acerca del escándalo que se levantó en el municipio de Almendralejo debido a la producción por parte de niños de entre 12 y 14 años de imágenes pornográficas generadas por inteligencia artificial en las que se representa la cara de menores de entre 11 y 17 años, ha generado cierto debate entre la opinión pública y distintos sectores del ámbito jurídico. Si bien estos hechos tuvieron lugar en el mes de septiembre de 2023, desde entonces ya se han registrado nuevos sucesos de menores víctimas del uso de esta tecnología con similares objetivos.[1]

Se trata una de entre las varias aplicaciones existentes que, a partir de imágenes de las menores con ropa tomadas de sus propias redes sociales, genera una serie de material pornográfico que posteriormente ha sido difundido entre los teléfonos móviles de los menores de este pueblo localizado en la provincia de Badajoz.

Enseguida, los medios de comunicación se han hecho eco de la preocupación social que en los últimos meses se ha generado ante el uso de este tipo de tecnología, siendo este el primero que ha trascendido en las redes en torno a la utilización de la imagen de menores para tales fines (anteriormente se habían dado casos, de los cual no se quiso dar noticia para evitar el efecto llamada).

Dada las dudas despertadas en torno a este tipo de supuestos, el objetivo del presente trabajo es determinar qué encaje tienen este tipo de conductas dentro de nuestro código penal, así como las problemáticas derivadas de ello. Igualmente, también se pretende hacer un breve acercamiento al perfil criminológico de estos presuntos delitos.

2. CALIFICACIÓN PENAL DEL HECHO.

2.1. Los deepfakes como producción y distribución de pornografía infantil.

Han sido varias las voces por parte de diversos juristas y medios de comunicación que han calificado este hecho como constitutivo de un delito de producción y distribución de pornografía infantil agravado del artículo 189.2.a) del código penal, con condena de prisión de hasta 9 años. No obstante, se debe tener en cuenta que dada la edad de cada uno de los sujetos activos en el caso concreto de Almendralejo, solo uno de ellos podría ser imputable por alcanzar los 14 años.

Sin embargo, la propia doctrina de la Fiscalía General del Estado excluye la posibilidad de castigar estos hechos bajo la aplicación del artículo 189.2 a) del código penal. No obstante, esto no es óbice para que los mismos puedan llegar a ser enmarcados dentro del artículo

1 https://elpais.com/espana/catalunya/2023-12-26/la-fiscalia-de-barcelona-investiga-fotos-sexuales-de- menores-manipuladas-con-inteligencia-artificial.html

189.1 b), siendo este un delito menos grave, debido a lo que sería producción y difusión de pseudo pornografía infantil.[2]

La producción de imágenes consistentes en un montaje realista entre la cara de un menor y el cuerpo de un adulto puede encajar en lo que ha sido generalmente catalogado por la doctrina como pseudo pornografía infantil. En palabras de la propia Fiscalía General del Estado, la pseudo pornografía ha sido clasificada en tres tipos: 1) imágenes de cuerpos digitalmente alteradas y sexualizadas, como la imagen de un niño en bañador al que se le quita dicha prenda con programas de ordenador; 2) imágenes separadas en una fotografía, como la mano de un niño sobreimpuesta a un pene de un adulto; 3) montajes de fotos, alguna de las cuales representa a un menor y otras tienen contenido sexual.[3]

El caso de Almendralejo podría enmarcarse en el primero de los tres tipos de pseudo pornografía infantil anteriormente mencionados. Este término fue introducido en el Código penal español a través de la reforma operada por LO 15/2003, de manera que el apartado séptimo del artículo 189 CP castigaba con la pena de prisión de tres meses a un año o multa de seis meses a dos años al que produjere, vendiere, distribuyere, exhibiere o facilitare por cualquier medio material pornográfico en el que no habiendo sido utilizados directamente menores o incapaces, se emplee su voz o imagen alterada o modificada

Con la reforma operada por LO 1/2015 se suprimió formalmente el tipo que castigaba la pseudo pornografía infantil. Sin embargo, la Doctrina de la Fiscalía General del Estado afirma que "*ello no supone la sobrevenida atipicidad de estas conductas, pues eventualmente podrán castigarse como pornografía virtual o técnica. Si se tipifican estas subespecies de pornografía, que no representan a menores reales, con más razón cabrá poder reaccionar contra la pseudo pornografía infantil, en la que se abusa de la imagen de un menor real. Ya el informe del Consejo Fiscal de 8 de enero de 2013 se pronunciaba en el sentido de que entender que "su supresión obedece a que tal material pornográfico debe reconducirse ahora a los supuestos de pornografía virtual que el Anteproyecto considera material pornográfico infantil relevante penalmente".* [4]

La supresión del antiguo tipo de pseudo pornografía infantil para fusionarlo con el castigo de la pornografía virtual (representaciones realistas pero ficticias de menores cuya distinción con la realidad presenta dificultades) y de la pornografía técnica (jóvenes que desconociéndose de si se tratan o no de menores son presentados como tales) genera inseguridad a la hora de definir qué tipo de imágenes pueden ser subsumidas en el presente tipo. La duda se incrementa si se tiene en cuenta la utilización de la palabra "eventual" por parte de la doctrina de la Fiscalía General del Estado a la hora de determinar si esta puede castigarse como pornografía virtual o técnica, dada la ausencia de jurisprudencia suficiente para poder delimitar en qué supuestos esta puede constituir delito.

2 Fiscalía General del Estado. Circular 2/2015, de 19 de junio, sobre los delitos de pornografía infantil tras la reforma operada por Ley Orgánica 1/2015.

3 Informe de Consejo Fiscal al anteproyecto de Ley Orgánica por la que se modifica la Ley Orgánica 10/1995, de 24 de noviembre, de Código Penal.

4 Fiscalía General del Estado. Circular 2/2015, de 19 de junio, sobre los delitos de pornografía infantil tras la reforma operada por Ley Orgánica 1/2015.

Del mismo modo, la DFGE excluye del concepto de pornografía infantil "los materiales que por su tosquedad revelen su condición de montaje". De nuevo, la falta de jurisprudencia provoca cierta inseguridad jurídica a la hora de determinar qué materiales se enmarcan en tal definición (una fotografía generada por la IA consistente en la cara de una menor de 11 años y el cuerpo de una mujer adulta genera pocas dudas acerca de su montaje, pero no porque este sea técnicamente defectuoso sino por la obvia diferencia en el desarrollo biológico entre ambos sujetos, por lo que de nuevo surge la incertidumbre acerca de si ello puede ser considerado como pornografía infantil).

No obstante, la propia doctrina de la Fiscalía General del Estado excluye explícitamente la posibilidad de poder considerar los agravantes del 189.2 CP en relación con los caso en los que el material pornográfico infantil fuera virtual o técnico, realizando una interpretación restrictiva de dichos agravantes, los cuales únicamente podrán ser considerados con respecto al 189.1. a) CP.

En cualquier caso, el castigo de esta clase de pornografía bajo el artículo 189.1. b) CP supondría enmarcar dentro del capítulo de los delitos relativos a la prostitución y a la explotación sexual y corrupción de menores un tipo de delito cuyo bien jurídico afectado no es la indemnidad ni la libertad sexual del menor, sino su imagen y derecho al honor. Ante esta situación cabe preguntarse si sería conveniente la redacción de un nuevo tipo que abarcase expresamente tales hechos y despejase tales dudas, pudiendo constituir un tipo propio o un tipo que castigue la producción de *deepfakes* sexuales y que incluya agravantes para los casos de menores o de contenido que revista especial importancia.

2.2. El fake porn cuando el sujeto pasivo es mayor de edad.

En relación con el final del epígrafe anterior, surge la cuestión acerca de si los *deepfakes* sexuales de sujetos mayores de edad son penalmente relevantes y bajo qué tipo podrían calificarse. A diferencia del caso de Almendralejo, este tipo de casos ya habían trascendido entre la prensa anteriormente. Múltiples estrellas del mundo del cine y de la música ya han sido previamente víctimas de la utilización de esta nueva tecnología. El último caso sonado en nuestro país ha sido el que la propia cantante Rosalía denunció en sus redes sociales.

Ante estos hechos, han sido varios los tipos a los que tanto la prensa como diversos juristas han señalado como posibilidad de enmarcar el fenómeno del *fake porn* bajo los mismos. Entre ellos, los más recurrentes han sido el delito de difusión, revelación o cesión de datos o imágenes captadas (art. 197.3º CP); el delito de sexting (197.7º CP); el delito de trato degradante (art. 173.1 CP párrafo 1º); y el delito de injurias (art 208 CP).

2.2.1. El delito de revelación de secretos (art 197.3 CP).

Algunos medios han señalado que la difusión de *deepfakes* de carácter sexual podría suponer un delito de revelación de secretos del artículo 197.3 CP. No obstante, a la hora de analizar el encaje de tales hechos en este artículo, es preciso tener en mente la descripción típica del mismo: *"Se impondrá la pena de prisión de dos a cinco años si se difunden, revelan o ceden a terceros los datos o hechos descubiertos o las imágenes captadas a que se refieren los números anteriores".*

El tipo del 197.3 CP se configura mediante remisión al del 197.1 CP y 197.2 CP, de manera que constituye un tipo agravado de ambos delitos. Este agravamiento se da cuando lo obtenido por medio de las conductas de los dos tipos anteriores pasa a ser difundido o cedido a terceros. Por ello, para determinar si la difusión de *fake porn* encaja dentro del tipo 197.3

CP habrá que comprobar previamente que esta conducta sea igualmente típica conforme al tipo básico del 197.1 CP o el tipo autónomo del 197.2 CP.

El artículo 197.1 CP castiga a quienes "*para descubrir los secretos o vulnerar la intimidad de otro, sin su consentimiento, se apodere de sus papeles, cartas, mensajes de correo electrónico o cualesquiera otros documentos o efectos personales, intercepte sus telecomunicaciones o utilice artificios técnicos de escucha, transmisión, grabación o reproducción del sonido o de la imagen, o de cualquier otra señal de comunicación".*

Todos los casos de *fake porn* que hasta el momento se han reportado ha sido elaborados bien a partir de la toma de fotografías del sujeto pasivo captadas en la calle, bien a partir de las imágenes que la propia víctima subía a sus redes sociales. Ello hace descartar la posibilidad de que esta conducta haya sido realizada mediante interceptación de las telecomunicaciones o utilización de artificios técnicos.

En cambio, en cuanto al apoderamiento, como indica JORGE BARRERO, de acuerdo con la jurisprudencia, este ha de entenderse en un sentido idéntico al de apropiación en los delitos contra el patrimonio. Ello supone que la toma fotografías o la descarga de las mismas de las redes sociales de la víctima no puede ser comprendida como "apoderamiento" en el sentido que se en el tipo del 197.1 CP. Este concepto de apoderamiento es igualmente aplicable al artículo 197.2 CP.[5]

Por tanto, la inaplicabilidad de ambos tipos a la conducta analizada, hace que tampoco esta pueda ser subsumible dentro del artículo 197.3 CP, descartando así el delito de revelación de secretos.

2.2.2. El delito de sexting (197.7 CP).

El artículo 197.7 tipifica el delito de sexting de la siguiente forma: *"Será castigado con una pena de prisión de tres meses a un año o multa de seis a doce meses el que, sin autorización de la persona*

5 Jorge Barreiro, A. El delito de descubrimiento y revelación de secretos en el código penal de 1995. Un an‡lisis del art'culo 197 del CP. Revista Jur'dica 6 (2002): 99-132

afectada, difunda, revele o ceda a terceros imágenes o grabaciones audiovisuales de aquélla que hubiera obtenido con su anuencia en un domicilio o en cualquier otro lugar fuera del alcance de la mirada de terceros, cuando la divulgación menoscabe gravemente la intimidad personal de esa persona".

En la redacción de este delito se pueden observar hasta cinco elementos requeridos para la concurrencia del los elementos objetivos del tipo:

a) Difundir, revelar o ceder a terceros imágenes o grabaciones audiovisuales del sujeto pasivo.

b) Actuar sin autorización del sujeto pasivo.

c) Que dichas imágenes o grabaciones audiovisuales hayan sido obtenidas con su anuencia

d) Que hayan sido obtenidas en un domicilio o cualquier otro lugar fuera del alcance de la mirada de terceros.

e) Que la divulgación menoscabe gravemente la intimidad personal del sujeto pasivo.

La Doctrina de la Fiscalía General del Estado, en su Circular 3/2017, aclara que *"Para que el precepto sea aplicable es necesario que la grabación objeto de difusión se haya llevado a efecto en un marco espacial de carácter reservado, circunstancia ésta que el tipo penal concreta en la exigencia de que se haya obtenido en un domicilio, o en un lugar fuera del alcance de la mirada de terceros (...) resulta esencial a efectos de asegurar el carácter íntimo de la imagen o grabación, el lugar de la realización o toma de la misma, que ha de tratarse de un espacio físico excluido, en ese momento, al conocimiento de terceros."*

Esta exigencia de que las imágenes sean captadas en un contexto de estricta intimidad y sustraído a la percepción de terceros ajenos parece descartar la tipicidad de sucesos como el de Almendralejo, en el que las imágenes fueron captadas de las redes sociales personales de las menores. No obstante, en caso de que dichas imágenes hubieran sido tomadas en dichos contextos de intimidad, cabe preguntarse si la utilización de las mismas para la producción de *fake porn* puede ser considerado como un delito de sexting.

En tal supuesto, la tipicidad de la conducta dependería de que la divulgación menoscabe gravemente la intimidad personal del sujeto pasivo. Ante esto, la Doctrina de la Fiscalía General del Estado indica que este es un elemento que debe valorarse caso a caso según las circunstancias concurrentes. No obstante, sí que menciona que la difusión de tal imagen deba provocar una seria injerencia en el ámbito personal de la intimidad de la víctima.

En este sentido, surge la duda acerca de si la imagen de la cara de una persona sin ningún tipo de identificación ni dato personal de esta pero fusionada con un cuerpo desnudo de un tercero a modo de montaje o *deepfake* puede considerarse que vulnera la intimidad del sujeto pasivo.

La ausencia de jurisprudencia relativa a esta casuística en nuestro país impide poder dar una respuesta contundente a tal cuestión.

Autores como JAREÑO LEAL defienden que el ataque a la intimidad es aquel que representa la verdad, de manera que la imagen íntima debe ser original, quedando fuera del tipo aquellas que simulan dicha intimidad a través de manipulaciones, como en este caso sería el uso de la IA. JAREÑO LEAL establece una analogía bastante clarificante, de forma que

del mismo modo que el abrir una carta ajena que ha sido simulada por un tercero no supone un acceso a la intimidad del autor con identidad simulada, la difusión de una imagen parcialmente simulada no puede suponer la vulneración del derecho a la intimidad.[6]

Pudiera plantearse que el hecho de que se difunda la imagen real de la cara del sujeto pasivo debe suponer un ataque a la intimidad. Sin embargo, la ausencia de ningún tipo de información que permita identificar los datos personales de la víctima debe hacernos descartar tal opción. Además, no se puede pasar por alto que la difusión del mero rostro del sujeto pasivo no es el hecho cuya supuesta antijuridicidad ha provocado que tales conductas hayan sido denunciadas. Es precisamente el cuerpo desnudo generado por la IA, dada su fusión con el rostro, lo que ha generado la situación de agravio para las menores de Almendralejo.

Resultan precisas entonces las palabras de Simó Soler, quien indica que debería añadirse un nuevo apartado al artículo 197 CP referente al modo de generación del contenido audiovisual, articulando un marco legal que prevea este tipo de conductas.

2.2.3. Los delitos de trato degradante y de injurias.

Mención aparte merecen el delito de trato degradante y el delito de injurias. La menor taxatividad en la redacción de ambos tipos los hace más proclives a poder incluir comportamientos como la producción y difusión de *deepfakes* cuyas víctimas son identificables. Si bien la difusión de imágenes es un medio válido para la comisión del delito de injurias, parece ser más idónea la consideración del delito de trato degradante a la hora de considerar como delictivos estos hechos. Ello principalmente en base a la gradación de la gravedad, el hecho de que la difusión de este tipo de contenido afectaría más al bien jurídico del honor que al de la imagen y sobre todo teniendo en cuenta la existencia ya de una sentencia del Tribunal Supremo que, si bien no condena al sujeto activo por difusión de *deepfakes*, lo hace por atentar contra la dignidad personal de una soldado compañera de armas al difundir la foto de una mujer desnuda que era muy parecida a ella.

Sin embargo, otros autores como JAREÑO LEAL han señalado que el delito de trato degradante requiere la esistencia de una relación de dominio entre el sujeto pasivo y el sujeto activo, de menra que el primero se encuentre totalmente sometido al segundo. Así, dado que el fin de tal conducta sería el de someter la voluntad del sujeto pasivo mediante un acto personal sobre la víctima, ello obstaculizaría el encaje del *fake porn* en este tipo.

Igualmente, no podemos obviar el hecho de que la mayoría de la doctrina entiende que tanto el delito de trato degradante como el de injurias son delitos de resultado, por lo que estos no encajaría del todo con el que debiera ser el resultado constitutivo de delito en la producción de *deepfakes* sexuale*s* (la producción y difusión del propio material en sí, sin necesidad

6 JAREÑO LEAL, Ángeles. El derecho a la imagen íntima y el Código penal. La calificación de los casos de elaboración y difusión del deepfake sexual. Revista Electrónica de Ciencia Penal y Criminología. 2024, núm. 26-09, pp. 1-37. http://criminet.ugr.es/recpc/26/recpc26-09.pdf

de que el sujeto pasivo adquiera conocimiento o se vea afectado personalmente por ello). De hecho, la propia sentencia recién citada recoge que la víctima padeció como consecuencia de la situación generada desasosiego e intranquilidad porque sentía que era objeto de comentarios y cuchicheos entre los demás miembros de la unidad y esto le hizo padecer algunos episodios de llanto. De esta forma, incluir el fenómeno del *fake porn* en el tipo penal del trato degradante podría suponer la atipicidad de aquellos *deepfakes* difundidos de los cuales las propias víctimas carecieran de conocimiento o cuyo contenido no les provocase un resultado de menoscabo en su integridad moral. Lo mismo ocurriría en el caso de aquellos *deepfakes* que incluyan la cara de un menor pero que puedan no ser considerados como pseudo pornografía infantil por los motivos anteriormente citados y sean difundidos entre usuarios de la red sin que la propia víctima tenga conocimiento (práctica habitual en lo relativo al delito de difusión de pornografía infantil propiamente dicha).

3. PROPUESTAS DE TIPIFICACIÓN.

En definitiva, resulta evidente que el desarrollo de la inteligencia artificial en lo relativo a la fabricación de este tipo de material supone un desafío para el que no está del todo claro si nuestro Derecho Penal actual está en condiciones de afrontar. Del mismo modo, ya son varios los Estados de EEUU que han tomado la iniciativa legislativa para tratar de adaptarse a este tipo de acontecimiento. Por ejemplo, el Estado de Virginia ha incluido la "distribución de imágenes reales o fotos falsamente creadas" dentro del delito de sexting, el cual se castiga con hasta 12 meses de prisión y 2.500 dólares de multa.[7] Por su parte, Luisiana ha convertido en delito crear o poseer a sabiendas un objeto generado por IA imagen o vídeo que represente a una persona menor de 18 años participando en un acto sexual, con penas de entre 5 y 20 años de cárcel y multa de 10.000 dólares.[8]

En lo relativo a España, el Grupo Parlamentario Sumar ha realizado una Proposición de Ley Orgánica en la que propone la tipificación de este tipo de conductas dentro de los delitos de injurias. En concreto, de la siguiente forma:

El Código Penal queda modificado como sigue:

Uno. Se crea un nuevo artículo 208 bis, dentro del capítulo III del título XII del Código Penal, nuevo precepto penal que tendrá la siguiente redacción:

«Artículo 208 bis.

Igualmente tendrá la consideración de injuria la acción que, sin autorización y con ánimo de menoscabar el honor, fama, dignidad o la propia estimación de una persona, recrease mediante sistemas automatizados, software, algoritmos o inteligencia artificial para la pública difusión su imagen corporal o audio de voz.»

7 https://www.eltiempo.com/tecnosfera/novedades-tecnologia/deepfakes-ya-son-un-crimen- de-porno-de-la-venganza-en-virginia-383092

8 https://www.heaven32.com/luisiana-proscribe-las-falsificaciones-sexuales-de-ninos-hechas-por-ia

Dos. Se modifica el artículo 211 del Código Penal, añadiendo un párrafo segundo, quedado redactado el artículo de la siguiente forma:

«Artículo 211.

La calumnia y la injuria se reputarán hechas con publicidad cuando se propaguen por medio de la imprenta, la radiodifusión o por cualquier otro medio de eficacia semejante. Salvo previa autorización expresa de la persona o personas afectadas, las simulaciones de imágenes, vídeos o audios de voz de estas generados a través de sistemas automatizados, software, algoritmos o mecanismos de inteligencia artificial que fueran difundidos a través de redes sociales serán consideradas como injurias hechas con publicidad.»

Esta propuesta de introducción de un nuevo artículo dentro de los delitos de injurias revela las dudas e inseguridad jurídica actual existente en torno al encaje de este tipo de conductas dentro del código penal. El encaje por parte del Grupo de Sumar de los *deepfake* dentro de las injurias puede resulta sin duda problemático. Si se propone la inclusión de estas conductas en el tipo que precisamente contempla una pena menor de todos los analizados, ello implica que se teme que dichas conductas sean atípicas conforme a nuestro código penal actual, pues en caso contrario no se optaría por expresar de tipificar explícitamente los *deepfakes* en un tipo que no contempla la pena de prisión.

Por otra parte, se estaría renunciando a contemplar los *deefakes* sexuales como delitos contra la intimidad. Así, se clasificarían todos los tipos de *deepfakes* bajo un mismo artículo, renunciando así a hacer una diferenciación entre, por ejemplo, los *deepfakes* de carácter político o electoral y los *deepfakes* de carácter sexual. Del mismo modo, ello no solucionaría el problema señalado anteriormente acerca del carácter del delito de injurias como un delito de resultado, que no penaría la mera producción de *fake porn* y su posterior difusión con desconocimiento de la víctima.

4. CONCLUSIÓN

La utilización de la Inteligencia Artificial para la producción de *deep fakes* de carácter sexual o *fake porn* es sin duda un fenómeno cuyas consecuencias jurídicas aún son inciertas, dada la ausencia de sentencias condenatorias por este tipo de actos. La novedad de este tipo de conductas y la inseguridad jurídico-penal existente con respecto a las mismas, hace que los intentos de legislar sobre ellas supongan intentos inciertos de intentar encajar lo que aún se desconoce si ya tiene encaje en nuestro código penal. Sin embargo, la espera a la existencia de sentencias que se pronuncien acerca de este tipo de sucesos supone el riesgo evidente de que dichas conductas puedan resultar impunes.

En definitiva, el propio carácter del Derecho frente a la tecnología hace que en este tipo de casos la ley siempre vaya por detrás de la realidad. Si bien los *deepfakes* podrían ser catalogados como delito de injurias, aun restan varios interrogantes en torno a los supuestos en los que los sujetos pasivos son menores de edad, así como en aquellos casos en los que la víctima no adquiere conocimiento de la existencia de tal material.

Deberán ser los tribunales los que en los próximos meses se pronuncien acerca de este fenómeno, de manera que sienten las bases necesarias para que el Legislativo pueda actuar en consecuencia y legislar con un objetivo definido en este ámbito.

5. BIBLIOGRAFÍA

BAUER, FELIPE. Los delitos de pornografía infantil como paradigma de moderno Derecho Penal. Sevilla, 2015.

DOMINGO JARAMILLO, C. (2021). Utilización del sistema de reconocimiento facial para preservar la seguridad ciudadana, El Criminalista Digital, 9, 20-37.

Fiscalía General del Estado. Circular 2/2015, de 19 de junio, sobre los delitos de pornografía infantil tras la reforma operada por Ley Orgánica 1/2015.

GONZÁLEZ VEGAS, A., PAZ-ARES MÁRQUEZ, I. Con acento jurídico: las implicaciones legales del Deepfake. 2021

Informe de Consejo Fiscal al anteproyecto de Ley Orgánica por la que se modifica la Ley Orgánica 10/1995, de 24 de noviembre, de Código Penal.

JAREÑO LEAL, ÁNGELES. El derecho a la imagen íntima y el Código penal. La calificación de los casos de elaboración y difusión del deepfake sexual. Revista Electrónica de Ciencia Penal y Criminología. 2024, núm. 26-09, pp. 1-37. http://criminet.ugr.es/recpc/26/recpc26-09.pdf

JORGE BARREIRO, A. El delito de descubrimiento y revelación de secretos en el código penal de 1995. Un análisis del artículo 197 del CP. Revista Jurídica 6 (2002): 99-132

Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal.

MARÍN DE ESPINOSA CEBALLOS, E. (Dir.). Lecciones de Derecho Penal. Parte Especial. Tirant lo Blanch. 5ª Edición, Valencia, 2021.

MICHELLE AZUAJE, P. Deepfakes, distorsión de la realidad y desafíos jurídicos. ¿Cómo debe responder el derecho cuando no todo es lo que parece?, 2023

MORENO-TORRES HERRERA, Mª. R (Directora), Zugaldía Espinar, J. M., Marín de Espinosa Ceballos, Mª E., Fernández Teruelo, J.G., Esquinas Valverde, P., García Amez, J., Morales Hernández, M.A. Lecciones de Derecho Penal. Parte General. Edit. Tirant lo Blanch. 2ª Edición, Valencia 2021.

https://www.elmundo.es/espana/2023/09/20/650b2ae7fc6c83256e8b45c5.html

https://www.elmundo.es/opinion/editorial/2023/09/21/650b47e1e9cf4af40b8b4594.html

https://www.elconfidencial.com/espana/2023-09-24/caso-almendralejo-menores-inteligencia-artificial_3740932/

https://www.xataka.com/privacidad/falsos-desnudos-menores-generados-ia-policia-investiga-almendralejo-primer-caso-masivo-espana

https://www.elmundo.es/madrid/2019/05/30/5cee6365fc6c83ae2a8b45a6.html

https://www.newtral.es/deepfakes-pornograficos-mujeres-ley-ia-almendralejo/20230920/

https://www.epe.es/es/igualdad/20230502/pornografia-deepfake-codigo-penal-violencia-sexual-86606820

https://www.ibericonnect.blog/2023/07/algunas-claves-del-castigo-penal-del-deepfake-de-naturaleza-sexual/

https://www.xataka.com/inteligencia-artificial/estados-unidos-estan-empezando-a-legislar-deepfakes-asi-esta-normativa-al-respecto-espana

https://www.20minutos.es/noticia/5174137/0/almendralejo-ley-inteligencia-artificial-ue-porno-futuro/

https://www.heaven32.com/luisiana-proscribe-las-falsificaciones-sexuales-de-ninos-hechas-por-ia/

https://www.eltiempo.com/tecnosfera/novedades-tecnologia/deepfakes-ya-son-un-crimen-de-porno-de-la-venganza-en-virginia-383092

Capítulo 10

LA REGULACIÓN DE LA INTELIGENCIA ARTIFICIAL EN EUROPA[1]

Miguel Ángel Presno Linera
Catedrático de Derecho constitucional de la Universidad de Oviedo
Anne Meuwese
Catedrática de Derecho público y gobernanza de la inteligencia artificial de la Universidad de Leiden

RESUMEN: Pocas semanas antes de finalizar la redacción de estas páginas se conocieron tanto el texto del Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (la llamada Ley de Inteligencia Artificial) como el del Convenio Marco del Consejo de Europa sobre inteligencia artificial, derechos humanos, democracia y Estado de Derecho. Ambas normas son el resultado presente de un proceso de debate legislativo, institucional, social, tecnológico y económico que se remonta a varios años atrás y que parte del convencimiento, como se dice en el Libro Blanco sobre la inteligencia artificial de la Comisión Europea, de que la inteligencia artificial "cambiará nuestras vidas, pues mejorará la atención sanitaria, aumentará la eficiencia de la agricultura, contribuirá a la mitigación del cambio climático y a la correspondiente adaptación, mejorará la eficiencia de los sistemas de producción a través de un mantenimiento predictivo, aumentará la seguridad de los europeos y nos aportará otros muchos cambios que de momento solo podemos intuir. Al mismo tiempo, la IA conlleva una serie de riesgos potenciales, como la opacidad en la toma de decisiones, la discriminación de género o de otro tipo, la intromisión en nuestras vidas privadas o su uso con fines delictivos". En las siguientes páginas analizaremos las respuestas jurídicas que se han dado en Europa al reto de regular la IA de forma no apocalíptica pero tampoco totalmente integrada.

PALABRAS CLAVE: Unión Europea, Consejo de Europa, derechos fundamentales, inteligencia artificial, regulación de la inteligencia artificial.

1 Este trabajo es uno de los resultados del Proyecto de investigación PID2022-136548NB-I00 *Los retos de la inteligencia artificial para el Estado social y democrático de Derecho*.

1. INTRODUCCIÓN: LAS INICIATIVAS PARA LA REGULACIÓN DE LA INTELIGENCIA ARTÍFICIAL EN EUROPA.

El 19 de octubre de 2017, el Consejo Europeo concluyó que, para construir con éxito una Europa digital, la Unión Europea (UE) necesitaba "concienciarse de la urgencia de hacer frente a las nuevas tendencias, lo que comprende cuestiones como la inteligencia artificial (IA) y las tecnologías de cadena de bloques, garantizando al mismo tiempo un elevado nivel de protección de los datos, así como los derechos digitales y las normas éticas. El Consejo Europeo ruega a la Comisión que, a principios de 2018, proponga un planteamiento europeo respecto de la inteligencia artificial y le pide que presente las iniciativas necesarias para reforzar las condiciones marco con el fin de que la UE pueda buscar nuevos mercados gracias a innovaciones radicales basadas en el riesgo y reafirmar el liderazgo de su industria".

Se comenzó a evidenciar así la preocupación de las instituciones de la UE por la regulación jurídica de la IA, aprovechando todo lo que supone en materia de innovación y desarrollo tecnológico y, al mismo tiempo, garantizando los derechos fundamentales y el Estado social y democrático de Derecho.

El 19 de febrero de 2020 la Comisión publicó el *Libro Blanco sobre la inteligencia artificial*, donde se respaldaba "un enfoque basado en la regulación y en la inversión, que tiene el doble objetivo de promover la adopción de la inteligencia artificial y de abordar los riesgos vinculados a determinados usos de esta nueva tecnología".

En el mes de octubre del mismo año, el Parlamento Europeo aprobó diversas resoluciones en materia de IA en el ámbito de la ética, la responsabilidad civil y los derechos de propiedad intelectual, a las que siguieron, ya en 2021, resoluciones sobre el uso de la IA y en los sectores educativo, cultural y audiovisual y en materia penal.

Especialmente importante fue la fecha del 21 de abril de 2021, cuando se conoció la Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión, elaborada por la Comisión.

Al respecto, el 6 de diciembre de 2022, el Consejo de la Unión Europea hizo pública su Orientación general de 25 de noviembre, donde señaló que "para garantizar que la definición de los sistemas de IA proporcione criterios suficientemente claros para distinguirlos de otros sistemas de software más clásicos, el texto transaccional restringe la definición a los sistemas desarrollados a través de estrategias de aprendizaje automático y basadas en la lógica y el conocimiento".

A continuación, cabe mencionar las enmiendas aprobadas por el Parlamento Europeo el 14 de junio de 2023, que incorporaron una nueva definición de sistema de IA –"un sistema basado en máquinas diseñado para funcionar con diversos niveles de autonomía y capaz, para objetivos explícitos o implícitos, de generar información de salida —como predicciones, recomendaciones o decisiones— que influya en entornos reales o virtuales"; también de lo que se entendió por un modelo fundacional –"un modelo de sistema de IA entrenado con un

gran volumen de datos, diseñado para producir información de salida de carácter general y capaz de adaptarse a una amplia variedad de tareas diferentes-", al tiempo que, entre otras cosas, se ampliaron los sistemas de IA prohibidos.

En diciembre de 2023 tuvieron lugar los «trílogos» entre las instituciones europeas (Parlamento, Comisión, Consejo) para limar las diferencias en cuestiones tan relevantes como la prohibición, o no, del uso de los sistemas de identificación biométrica remota «en tiempo real» en espacios de acceso público o el alcance de la regulación de los entonces denominados "modelos fundacionales", que pasaron a llamarse "modelos de IA de uso general".

Todo este proceso desembocó, finalmente, en la Resolución legislativa del Parlamento Europeo, de 13 de marzo de 2024, sobre la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión (en adelante, RIA).

Por lo que respecta al Consejo de Europa, destaca, en primer lugar, el trabajo de investigación sobre algoritmos y derechos humanos, de 2017, según el cual la IA afectará a un gran número, sino a la práctica totalidad, de nuestros derechos fundamentales; así, al derecho a la libertad personal y, muy relacionado con él, al derecho a un juicio justo y a la tutela de los tribunales; en segundo lugar, a los derechos de las personas en su dimensión más privada, como el derecho a la intimidad y a la protección de datos; en tercer lugar, a los derechos vinculados a la dimensión pública y relacional de las personas, como las libertades de expresión, información, creación artística e investigación pero también a las libertades de reunión y asociación, tanto en el plano meramente ciudadano como en lo que se refiere, por ejemplo, al ámbito laboral (libertad sindical, derecho de huelga); en cuarto lugar, y a su vez vinculado a muchos otros derechos, al de no sufrir discriminación por raza, género, edad, orientación sexual…; en quinto lugar, a los derechos dependientes del acceso a los servicios públicos (educación, sanidad…) y, en general, a los derechos sociales (prestaciones por desempleo, enfermedad, jubilación…); finalmente, y por no extendernos mucho más, al derecho a intervenir en procesos participativos de índole política (elecciones, referendos, iniciativas legislativas populares…) y en, general, a las libertades en el ámbito ideológico (de pensamiento, conciencia y religión).

En segundo lugar, la Asamblea Parlamentaria del Consejo de Europa aprobó un conjunto de principios éticos básicos que deberían respetarse al elaborar y establecer aplicaciones de IA, incluida la transparencia, la justicia y la equidad, la responsabilidad humana de la toma de decisiones, la seguridad, la privacidad y la protección de datos.

En tercer lugar, el Comité de Ministros adoptó un enfoque transversal de la IA en los diversos sectores del Consejo de Europa, estableciendo el Comité sobre Inteligencia Artificial (CAI) y encomendándole la elaboración de un Convenio [marco] jurídicamente vinculante sobre el desarrollo, diseño y aplicación de sistemas de IA, basado en las normas del Consejo de Europa en materia de derechos humanos, democracia y estado de derecho, sobre la base de estos principios fundamentales.

El Comité de Ministros también decidió permitir la inclusión en las negociaciones de la Unión Europea y de los Estados no europeos interesados que compartan los valores y ob-

jetivos del Consejo de Europa; a saber, Argentina, Australia, Canadá, Costa Rica, la Santa Sede, Israel, Japón, México, Perú, los Estados Unidos de América y Uruguay. El Consejo de Europa también hizo partícipes a actores no estatales en las negociaciones: un total de 68 representantes de la sociedad civil y de la industria lo hicieron en calidad de observadores, interviniendo junto con Estados y representantes de otras organizaciones internacionales, como la OSCE, la OCDE, la UNESCO y los órganos y comités pertinentes del Consejo de Europa. La Unión Europea también participó en las negociaciones representada por la Comisión Europea, incluyendo en su delegación también a representantes de la Agencia de los Derechos Fundamentales de la Unión Europea (FRA) y del Supervisor Europeo de Protección de Datos (SEPD).

El resultado final fue la conclusión, el 17 de mayo de 2024, del Convenio Marco sobre inteligencia artificial, derechos humanos, democracia y Estado de Derecho (en adelante el CIA).

2. ¿DE QUÉ SE HABLA CUANDO SE HABLA DE INTELIGENCIA ARTIFICIAL EN EUROPA?

La dificultad de ofrecer una definición "acabada" de la IA se evidencia leyendo las diferentes versiones que se han ido ofreciendo durante el proceso de aprobación del RIA: en el texto que presentó la Comisión en 2021 se entendía como "el software que se desarrolla empleando una o varias de técnicas y estrategias que figuran en el Anexo I y que puede, para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa".

Tras las enmiendas aprobadas por el Parlamento Europeo en 2023, se definió como "un sistema basado en máquinas diseñado para funcionar con diversos niveles de autonomía y capaz, para objetivos explícitos o implícitos, de generar información de salida —como predicciones, recomendaciones o decisiones— que influya en entornos reales o virtuales".

En el texto definitivo se entiende como un sistema basado en máquinas diseñado para funcionar con distintos niveles de autonomía, que puede mostrar capacidad de adaptación tras su despliegue y que, para objetivos explícitos o implícitos, infiere, a partir de las entradas que recibe, salidas tales como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales (artículo 3).

En el Preámbulo se explica que una característica clave de los sistemas de IA es su capacidad para inferir. Las técnicas que permiten la inferencia incluyen enfoques de aprendizaje automático y enfoques basados en la lógica y el conocimiento que infieren a partir del conocimiento codificado o la representación simbólica de la tarea que debe resolverse. La capacidad de un sistema de IA para inferir va más allá del procesamiento básico de datos, permitiendo el aprendizaje, el razonamiento o el modelado. Los sistemas de IA están diseñados para funcionar con distintos niveles de autonomía, lo que significa que tienen cierto grado

de independencia de las acciones de la intervención humana y capacidades para funcionar sin intervención humana. La capacidad de adaptación que puede mostrar un sistema de IA tras su despliegue se refiere a las capacidades de autoaprendizaje, que permiten al sistema cambiar mientras se utiliza.

Esta definición es muy similar a la del CIA, en cuyo artículo 2 se dice que se entenderá por «sistema de IA» un sistema basado en máquinas que, con objetivos explícitos o implícitos, infiere, a partir de los datos que recibe, cómo generar resultados, como predicciones, contenidos, recomendaciones o decisiones que puedan influir en entornos físicos o virtuales. Los diferentes sistemas de IA varían en sus niveles de autonomía y adaptabilidad después de la implementación".

3. LOS PRINCIPIOS QUE INSPIRAN LA REGULACIÓN DE LA INTELIGENCIA ARTIFICIAL EN EUROPA.

El Grupo independiente de expertos de alto nivel sobre IA creado por la Comisión Europea hizo públicos en abril de 2019 una serie de principios con el objeto de contribuir a garantizar la fiabilidad y el fundamento ético de la IA: acción y supervisión humanas; solidez técnica y seguridad; gestión de la privacidad y de los datos; transparencia; diversidad, no discriminación y equidad; bienestar social y ambiental, y rendición de cuentas.

Y, como se recuerda en el considerando 27 del RIA, por acción y supervisión humanas se entiende que los sistemas de IA se desarrollan y utilizan como una herramienta al servicio de las personas, que respeta la dignidad humana y la autonomía personal, y que funciona de manera que pueda ser controlada y vigilada adecuadamente por seres humanos; por solidez técnica y seguridad se entiende que los sistemas de IA se desarrollan y utilizan de manera que sean sólidos en caso de problemas y resilientes frente a los intentos de alterar el uso o el funcionamiento del sistema de IA para permitir su uso ilícito por terceros y reducir al mínimo los daños no deseados; por gestión de la privacidad y de los datos se entiende que los sistemas de IA se desarrollan y utilizan de conformidad con normas en materia de protección de la intimidad y de los datos, al tiempo que tratan datos que cumplen normas estrictas en términos de calidad e integridad; por transparencia se entiende que los sistemas de IA se desarrollan y utilizan de un modo que permita una trazabilidad y explicabilidad adecuadas, y que, al mismo tiempo, haga que las personas sean conscientes de que se comunican o interactúan con un sistema de IA e informe debidamente a los responsables del despliegue acerca de las capacidades y limitaciones de dicho sistema de IA y a las personas afectadas acerca de sus derechos; por diversidad, no discriminación y equidad se entiende que los sistemas de IA se desarrollan y utilizan de un modo que incluya a diversos agentes y promueve la igualdad de acceso, la igualdad de género y la diversidad cultural, al tiempo que se evitan los efectos discriminatorios y los sesgos injustos prohibidos por el Derecho nacional o de la Unión; por bienestar social y ambiental se entiende que los sistemas de IA se desarrollan y utilizan de manera sostenible y respetuosa con el medio ambiente, así como en beneficio de todos los

seres humanos, al tiempo que se supervisan y evalúan los efectos a largo plazo en las personas, la sociedad y la democracia.

Estos principios se concretan a lo largo del articulado del RIA y, en una línea similar, aunque de manera más genérica, el CIA prevé, entre las obligaciones generales para los sistemas de IA, las de transparencia y supervisión; las de igualdad y no discriminación; las de privacidad y protección de datos personales; la de la fiabilidad y la de la transparencia.

4. UN ENFOQUE DE LA REGULACIÓN DE LA INTELIGENCIA ARTIFICIAL BASADO EN LOS RIESGOS.

La regulación de la IA, tal y como se concibe en Europa, aunque no solo en este espacio jurídico, requiere la aplicación de un enfoque basado en los riesgos, que adapte las normas y su contenido a la intensidad y el alcance de los riesgos que puedan generar los sistemas de IA. Se trata de una de las concreciones del bien conocido «principio de precaución», que guía la actuación de la Unión Europea.

El RIA define el riesgo como la combinación de la probabilidad de que se produzca un perjuicio y la gravedad de dicho perjuicio (artículo 3.2) y, como consecuencia, en algunos casos el riesgo resultante será inaceptable y eso conducirá a la prohibición de los sistemas que lo generen. Otros sistemas se caracterizan de alto riesgo porque son capaces de causar un perjuicio a la salud, la seguridad o los derechos fundamentales de las personas físicas; por este motivo se les someterá (artículo 9) a un proceso iterativo continuo, planificado y ejecutado durante todo el ciclo de vida, que requerirá revisiones y actualizaciones sistemáticas periódicas. Constará de las siguientes etapas: a) la determinación y el análisis de los riesgos conocidos y previsibles que el sistema de IA pueda plantear para la salud, la seguridad o los derechos fundamentales cuando se utilice de conformidad con su finalidad prevista; b) la estimación y la evaluación de los riesgos que podrían surgir cuando el sistema de IA se utilice de conformidad con su finalidad prevista y cuando se le dé un uso indebido razonablemente previsible; c) la evaluación de otros riesgos que podrían surgir, a partir del análisis de los datos recogidos con el sistema de vigilancia post-comercialización; d) la adopción de medidas adecuadas y específicas de gestión de riesgos diseñadas para hacer frente a los riesgos detectados…

Con vistas a eliminar o reducir los riesgos asociados a la utilización del sistema de IA de alto riesgo, se tendrán en cuenta los conocimientos técnicos, la experiencia, la educación y la formación que se espera que posea el responsable del despliegue, así como el contexto en el que está previsto que se utilice el sistema.

También el CIA contempla un marco jurídico de gestión de riesgos e impactos (artículo 16), imponiendo a los Estados parte medidas para la identificación, evaluación, prevención y mitigación de los riesgos planteados por los sistemas de inteligencia artificial, teniendo en cuenta los impactos reales y potenciales en los derechos humanos, la democracia y el Estado de Derecho. Dichas medidas se graduarán y diferenciarán, según proceda, y deberán tener en cuenta el contexto y el uso previsto de los sistemas de IA, en particular en lo que respecta

a los riesgos para los derechos humanos, la democracia y el Estado de Derecho; también la gravedad y la probabilidad de los posibles impactos; considerarán, cuando proceda, las perspectivas de las partes interesadas, en particular de las personas cuyos derechos puedan verse afectados; incluirán documentación de los riesgos, los impactos reales y potenciales, y el enfoque de gestión de riesgos; exigirán que se sometan a ensayo los sistemas de IA antes de ponerlos a disposición para su primer uso y cuando se modifiquen significativamente. Finalmente, cada Parte evaluará la necesidad de una moratoria o prohibición u otras medidas adecuadas con respecto a determinados usos de los sistemas de IA cuando considere que dichos usos son incompatibles con el respeto de los derechos humanos, el funcionamiento de la democracia o el Estado de Derecho.

5. LAS DIFERENCIAS ESTRUCTURALES ENTRE EL REGLAMENTO DE LA UNIÓN EUROPEA Y EL CONVENIO MARCO DEL CONSEJO DE EUROPA.

En primer lugar, el ámbito de aplicación del RIA es, lógicamente, el territorio de la UE, mientras que el CIA nace con una cierta vocación de globalidad y prevé que estará abierto a la firma de los Estados miembros del Consejo de Europa, de los Estados no miembros que hayan participado en su elaboración y de la UE.

En segundo lugar, el RIA será obligatorio en todos sus elementos y directamente aplicable en cada Estado miembro; el CIA será obligatorio en la medida en los Estados se incorporen a él.

En tercer término, el RIA incluye una normativa extensa (113 artículos y XIII Anexos) y muy prolija, con preceptos muy detallados, mientras que el CIA se compone de un articulado mucho más reducido (36 preceptos) y menos detallado. El propio CIA destaca su "carácter marco", que podrá complementarse con otros instrumentos para abordar cuestiones específicas relacionadas con las actividades dentro del ciclo de vida de los sistemas de inteligencia artificial.

En cuarto lugar, el RIA, en términos generales, contiene más reglas, es decir, incluye comportamientos precisos de lo que puede, o no, hacerse en materia de IA; el CIA adopta una configuración más principialista, esto es, contiene mandatos de optimización que pueden ser cumplidos en diferente grado.

En quinto lugar, y también con carácter general, el RIA impone obligaciones de medios y de resultado mientras que el CIA incluye, esencialmente, obligaciones de resultado, dejando a los Estados la concreción de las medidas adecuadas para alcanzarlos.

Finalmente, el RIA contiene un sistema de sanciones mientras que el CIA se limita a disponer que cada Parte establecerá o designará uno o más mecanismos eficaces para supervisar el cumplimiento de las obligaciones establecidas.

6. LOS POSIBLES "EFECTO BRUSELAS" Y "EFECTO ESTRASBURGO" DE LA REGULACIÓN EUROPEA DE LA INTELIGENCIA ARTIFICIAL.

En un conocido artículo publicado en 2012, que adoptó formato de libro en 2020, la profesora Anu Bradford explicó cómo y por qué las normas y reglamentos "de Bruselas" han penetrado en muchos aspectos de la vida económica dentro y fuera de Europa a través del proceso de "globalización normativa unilateral". La potencia del mercado interior de la UE, unido a unas instituciones reguladoras con buena reputación, obliga a las empresas extranjeras que quieran participar en ese mercado a adaptar su conducta o su producción a las normas de la UE, que a menudo son las más estrictas; la alternativa es la renuncia a ese mercado. Explica Bradford que las empresas multinacionales suelen tener un incentivo para estandarizar su producción a escala mundial y adherirse a una única norma. Esto convierte a la norma de la UE en una norma mundial: es el "efecto Bruselas de facto". Y, una vez que estas empresas orientadas a la exportación hayan ajustado sus prácticas empresariales para cumplir las estrictas normas de la UE, a menudo tienen el incentivo de presionar a sus gobiernos para que adopten esas mismas normas en un esfuerzo por igualar las condiciones frente a las empresas nacionales no exportadoras: el "efecto Bruselas de iure".

Pues bien, cabría pensar que la regulación europea de la IA podría generar, en la línea de lo que ha ocurrido en ámbitos como la vida privada y la protección de datos, una exportación del contenido de esa nueva normativa a otros Estados, un "efecto Bruselas" sobre la regulación de la IA. Sin embargo, la propia Bradford se ha mostrado escéptica al respecto en su último trabajo -*Digital Empires: The Global Battle to Global Battle to Regulate Technology*-, de 2023, recordando que Estados Unidos sigue siendo un modelo basado en el mercado abierto, China un modelo de centralismo estatal y la Unión Europea sigue apostando por la regulación.

Ahora bien, Estados Unidos también ha optado por aprobar normas que regulen la IA, aunque no sea con la misma intensidad que en la Unión Europeas; así, el 30 de octubre de 2023 el presidente Biden emitió la *Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*, donde se proclama que el Gobierno Federal tratará de promover principios y acciones responsables de seguridad y protección de la IA con otras naciones, "incluidos nuestros competidores". Además, y en la línea de la UE, en esa Orden se define la IA como un sistema basado en máquinas que puede, para un conjunto dado de objetivos definidos por el ser humano, hacer predicciones, recomendaciones o tomar decisiones que influyan en entornos reales o virtuales. Y se enuncian los ocho principios que deben guiar el desarrollo de la IA: la seguridad de los sistemas, la innovación responsable, el compromiso con los trabajadores, avance en igualdad y derechos, protección de los consumidores, protección de la intimidad, gestión de los riesgos y uso responsable de la IA, búsqueda del liderazgo social, económico y tecnológico.

En definitiva, y aunque en el caso de la regulación de la IA el impacto del "efecto Bruselas" pueda ser menor que en otros ámbitos, no parece en absoluto, por lo que está ocurriendo en otros espacios jurídicos, que esta propuesta vaya a tener repercusiones únicamente hacia dentro de la Unión.

Pero, además, de este "efecto Bruselas" también se podría hablar de un posible "efecto Estrasburgo", en este caso de una manera voluntaria, pues, como ya hemos visto, en el proceso de elaboración del Convenio Marco se incluyó en las negociaciones a la Unión Europea y a Estados no europeos, además de a representantes de otras organizaciones internacionales.

Además, en el artículo 25.1 del CIA se alienta a las Partes a que presten asistencia a los Estados que no sean Partes en la Convención para que actúen de conformidad con sus disposiciones y pasen a ser Partes en ella. Y en el artículo 31.1 se prevé que, después de la entrada en vigor del Convenio, el Comité de Ministros del Consejo de Europa podrá, previa consulta a las Partes y obteniendo su consentimiento unánime, invitar a cualquier Estado no miembro del Consejo de Europa que no haya participado en la elaboración del Convenio a adherirse. En suma, el Convenio Marco del Consejo de Europa se abre a la posibilidad de ser una norma global; veremos hasta dónde llega.

7. CONCLUSIONES

1.- Regular la IA no es una opción, es una necesidad dado el desarrollo que han alcanzado estos sistemas y los riesgos que su uso implica para los derechos fundamentales.

2.- Sería positiva una regulación lo más global posible de la IA, al menos con carácter principial, aunque no parece probable que compartan el enfoque regulatorio sistemas jurídicos tan diversos como el europeo, el norteamericano o el chino.

3.- La regulación de la IA debe otorgar protección a los derechos fundamentales afectados sin que ello suponga paralizar la innovación y el desarrollo tecnológicos. No obstante, en caso de conflicto, y en coherencia con el principio de precaución de la EU, debe prevalecer la tutela de los derechos.

4.- Las medidas que limiten el desarrollo y el uso de los sistemas de IA deben ser proporcionales al riesgo de cada sistema.

5.- Los "principios europeos" – intervención y vigilancia humanas, solidez y seguridad, gobernanza de los datos, transparencia, no discriminación y equidad, bienestar social y medioambiental- parecen razonables aunque son también mejorables.

6.- Es muy difícil para el Derecho regular algo tan disruptivo y dinámico como la IA pero es necesario afrontar este reto y hacerlo de forma coordinada e interdisciplinar no solo dentro del propio Derecho sino en directa relación con la Informática, la Medicina, la Lingüística, la Educación, etc.

8. BIBLIOGRAFÍA

BARRIO ANDRÉS, Moisés: "Inteligencia artificial: origen, concepto, mito y realidad", *El Cronista del Estado Social y Democrático de Derecho (Ejemplar dedicado a Inteligencia artificial y derecho)* (2022), pp. 14-21.

BARRIO ANDRÉS, Moisés: "Consideraciones sobre el ámbito extraterritorial del Reglamento Europeo de Inteligencia Artificial", *Derecho Digital e Innovación* (2024), nº 20.

BRADFORD, Anu: *The Brussels Effect: How the European Union Rules the World*, Oxford: Oxford University Press, 2020.

BRADFORD, Anu: *Digital Empires. The Global Battle to Regulate Technology*, Oxford: Oxford University Press, 2023.

COTINO HUESO, Lorenzo: "Un análisis crítico constructivo de la Propuesta de Reglamento de la Unión Europea por el que se establecen normas armonizadas sobre la Inteligencia Artificial (Artificial Intelligence Act)", *Diario La Ley*, 2 de julio, 2021.

GONZÁLEZ CABANES, Francisco y DÍAZ DÍAZ, N. ¿Qué es la Inteligencia Artificial? En Gamero Casado, Eduardo/Pérez Guerrero, Francisco (dirs.): *Inteligencia artificial y sector público: retos, límites y medios*, Valencia: Tirant lo Blanch, 2023 pp. 37-72, 2023.

PRESNO LINERA, Miguel Ángel: *Derechos fundamentales e inteligencia artificial*, Madrid: Marcial Pons, 2022.

PRESNO LINERA, Miguel Ángel: "La propuesta de "Ley de Inteligencia Artificial" europea", *Revista de las Cortes Generales* (2023), nº 116, pp. 81-133.

Capítulo 11

ANÁLISIS DEL CONVENIO MARCO SOBRE INTELIGENCIA ARTIFICIAL DEL CONSEJO DE EUROPA

Ana Gascón Marcén[1]

Profesora Permanente Laboral
Universidad de Zaragoza, Facultad de Derecho

RESUMEN: Esta comunicación analiza el nuevo Convenio Marco sobre Inteligencia Artificial y Derechos Humanos, Democracia y Estado de Derecho del Consejo de Europa. Se subraya su importancia como primer instrumento internacional jurídicamente vinculante sobre la materia y se explica su relación con el Reglamento de Inteligencia Artificial de la Unión Europea, así como el papel de ésta en sus negociaciones. No obstante, también se critican algunos de sus puntos débiles. Se trata de un avance bienvenido y necesario, pero el balance es agridulce, al tratarse de una oportunidad pérdida, dado que se ha sacrificado parcialmente el contenido material del mismo por un mayor consenso. En varios casos, se ha buscado contentar a la Unión Europea o a los Estados Unidos llevando a que el texto sea un mínimo común denominador con un limitado ámbito de aplicación.

PALABRAS CLAVE: Inteligencia Artificial, Derechos Humanos, Estado de Derecho, Consejo de Europa, Reglamento de Inteligencia Artificial

1 Miembro del equipo de investigación del proyecto «Hacia una transición digital centrada en la persona en la Unión Europea». Esta publicación es parte del proyecto TED2021-129307A-I00, financiado por MICIU/AEI/10.13039/501100011033

1. INTRODUCCIÓN

El Consejo de Europa es líder en la regulación internacional digital con el Convenio sobre Ciberdelincuencia de 2001 o el Convenio para la protección de las personas respecto al tratamiento automatizado de datos personales de 1981, ambos ratificados por un gran número de Estados no europeos.[2] Además, es un referente en protección de los derechos humanos, estado de derecho y democracia, principios que se ven reinterpretados y amenazados por la inteligencia artificial (IA).[3]

Esta organización creó en 2018 el *Committee of experts on Human Rights Dimensions of automated data processing and artificial intelligence*[4]; en 2019, el *Ad Hoc Committee on Artificial Intelligence* (CAHAI)*;* [5] y, en 2022, empezó a funcionar finalmente el *Committee on Artificial Intelligence* (CAI) encargado de negociar el texto del nuevo convenio sobre la materia.

El 17 de mayo de 2024, el Comité de Ministros del Consejo de Europa adoptó el Convenio Marco sobre Inteligencia Artificial y Derechos Humanos, Democracia y Estado de Derecho, que se abrió formalmente a la firma en Vilna el 5 de septiembre de este año.[6] Andorra,

2 GSTREIN, Oskar Josef, "The Council of Europe as an Actor in the Digital Age: Past Achievements", *Future Perspectives. Festschrift der Mitarbeiter* Innen und Doktorand* Innen zum*, 60, 2019, pp. 77-90.

3 Véase COMMITTEE OF EXPERTS ON INTERNET INTERMEDIARIES (MSI-NET), *Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications,* DGI(2017)12; LESLIE, David, et al. *Artificial intelligence, human rights, democracy, and the rule of law: a primer*. Council of Europe and The Alan Turing Institute, 2021 y GASCÓN MARCÉN, Ana, "Derechos humanos e inteligencia artificial", en: *Setenta años de Constitución Italiana y cuarenta años de Constitución Española*, Vol. 5, BOE y Centro de Estudios Políticos y Constitucionales, 2020, pp. 335-350.

4 Véase *Recommendation CM/2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems*. Además, en 2018 la Comisión Europea para la Eficiencia de la Justicia adoptó la *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment*.

5 Véase CAHAI, *Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law*, CAHAI(2021)09rev. Disponible en: https://rm.coe.int/cahai-2021-09rev-elements/1680a6d90d (última consulta: 17/06/2024). MANTELERO y FANUCCI analizaron este proceso destacando las tensiones subyacentes, los diferentes puntos de vista, las metodologías de consulta y las dinámicas que hacen del diálogo sobre la regulación de la IA un laboratorio extremadamente interesante para el derecho de los derechos humanos. En su opinión, las ambiciones reales que caracterizan el debate sobre la regulación de la IA y las altas expectativas iniciales de abordar cuestiones cruciales de derechos humanos podían verse frustradas y limitadas a un enfoque general basado en el riesgo, así como a una regulación minimalista más centrada en la industria y los beneficios económicos que en los derechos humanos. Véase MANTELERO, Alessandro; FANUCCI, Francesca, "Great Ambitions. The International Debate on AI Regulation and Human Rights in the Prism of the Council of Europe's CAHAI", *European Yearbook on Human Rights* 2022. Cambridge, Intersentia, pp. 225-252.

6 Disponible en: https://rm.coe.int/1680afae3c con su Informe explicativo en https://rm.coe.int/1680afae67 (última consulta: 17/06/2024). Después habrá que esperar probablemente varios años para que entre en vigor con su ratificación por cinco Estados. Como elemento de comparación, el último convenio adoptado en esta organización fue el Segundo Protocolo adicional al Convenio sobre la Ciberdelincuencia, relativo a la cooperación reforzada y la revelación de pruebas electrónicas que se abrió a la firma el 12 de mayo de 2022 y que también requería cinco ratificaciones para entrar en vigor. El mismo no ha entrado en vigor todavía más de dos años después, porque a 17 de junio de 2024, aunque ya cuenta con 44 firmas, únicamente dos Estados lo han ratificado.

Georgia, Islandia, Israel, Noruega, Moldavia, San Marino, Reino Unido, Estados Unidos de América y la Unión Europea (UE) lo firmaron ese mismo día.

Lo que hace especialmente relevante este Convenio es que se trata del primer tratado internacional sobre esa materia. Hasta ahora las iniciativas en este sector habían sido de *soft-law* (como los Principios de la OCDE sobre Inteligencia Artificial[7] o la Recomendación de la UNESCO sobre la ética de la inteligencia artificial[8]). Sin embargo, este Convenio será jurídicamente vinculante para los Estados que lo ratifiquen.

Su vocación es crear un marco regulador no sólo para sus 46 Estados miembros, sino ir más allá de las fronteras europeas. Han participado en su negociación otros once Estados: Argentina, Australia, Canadá, Costa Rica, Estados Unidos, la Santa Sede, Israel, Japón, México, Perú y Uruguay.[9] Estos Estados podrán ratificar el tratado, pero además también podrá hacerlo cualquier otro Estado del mundo que reciba una invitación del Comité de Ministros del Consejo de Europa.

A continuación, se analiza el Convenio, su relación con el Reglamento de Inteligencia Artificial de la Unión Europea (UE) y algunas de las principales críticas que ha recibido.

2. CONTENIDO DEL CONVENIO

El Convenio utiliza la definición de IA de la OCDE y recoge en su Capítulo II una serie de obligaciones generales relativas a la protección de los derechos humanos y la integridad de los procesos democráticos y respeto del estado de derecho. El Capítulo III regula los principios relacionados con las actividades dentro del ciclo de vida de los sistemas de IA: dignidad humana y autonomía individual; transparencia y supervisión; rendición de cuentas y responsabilidad; igualdad y no discriminación; privacidad y protección de datos personales; confiabilidad e innovación segura. Hay un brevísimo Capítulo IV sobre remedios.

7 El Consejo de la OCDE adoptó en 2019 una Recomendación con una serie de principios sobre la IA que se actualizó en mayo de 2024. La UE, el Consejo de Europa, los Estados Unidos, las Naciones Unidas y otros Estados utilizan la definición de la OCDE de un sistema de IA y su ciclo de vida que figura en esta recomendación en sus marcos legislativos, estrategias y orientaciones. Además, la OCDE ha creado un Observatorio de la IA para centralizar información sobre la materia. Véase: https://oecd.ai/en/ai-principles

8 Se considera el "primer marco normativo universal sobre ética de la IA". Si bien no es jurídicamente vinculante, su valor estriba en que muestra el consenso de la comunidad internacional en la materia al ser adoptado por los 193 Estados miembros de la UNESCO en noviembre de 2021. Disponible en: https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia-artificial (última consulta: 17/06/2024).

9 Además, han participado en las reuniones otras organizaciones internacionales y regionales como la UE (representada por la Comisión Europea), la ONU (en particular la UNESCO), la OCDE o la OSCE; representantes del sector privado, incluidas empresas y asociaciones; y representantes de la sociedad civil, instituciones académicas y de investigación admitidos como observadores, pero éstos últimos no han participado realmente en la redacción.

El Capítulo V se dedica a la evaluación y mitigación de riesgos e impactos adversos. Cada Parte adoptará medidas para la identificación, evaluación, prevención y mitigación de los riesgos planteados por los sistemas de IA, considerando los impactos reales y potenciales sobre los derechos humanos, la democracia y el estado de derecho. Dichas medidas serán graduadas y diferenciadas, con un enfoque basado en el riesgo, porque tendrán en cuenta el contexto y el uso previsto de los sistemas de IA, y la gravedad y probabilidad de posibles impactos. No se prohíbe el uso de ningún tipo de sistema de IA, sino que cada Parte evaluará la necesidad de una moratoria o prohibición u otras medidas apropiadas con respecto a ciertos usos de los sistemas de IA cuando considere que dichos usos son incompatibles con el respeto de los derechos humanos, el funcionamiento de la democracia o el estado de derecho.

El Capítulo VI se consagra a la aplicación del Convenio, con cuestiones transversales tales como la no discriminación, los derechos de las personas con discapacidad y de los niños, las consultas públicas, la alfabetización y competencias digitales, la salvaguardia de los derechos humanos y la posibilidad de regular una protección más amplia a nivel nacional.

Se ha incluido un Capítulo VII que establece mecanismos de seguimiento y cooperación, con la creación de una Conferencia de las Partes, la obligación de informar periódicamente sobre las actividades realizadas para dar efecto al Convenio, un artículo sobre cooperación internacional y la obligación de crear o designar mecanismos eficaces para supervisar el cumplimiento del Convenio a nivel nacional. La Conferencia de las Partes ofrecerá una interpretación dinámica de las disposiciones del Convenio, teniendo en cuenta las novedades que se puedan dar en el sector de la IA, y, además, la presión por pares fomentará la implementación del Convenio.

3. RELACIÓN CON EL REGLAMENTO DE IA DE LA UE

La negociación y finalización del Convenio ha quedado opacada por la negociación y adopción en paralelo del Reglamento de la UE por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento de Inteligencia Artificial o RIA, conocido como *AI Act)*. Entre ambas normas existe un solapamiento. Si bien son de naturaleza diferente, dado que el Convenio sólo obligará a los Estados que lo ratifiquen, pero éstos pueden ir mucho más allá de las fronteras europeas. Mientras que el RIA sólo obligará a los Estados miembros de la UE,[10] pero en éstos tendrá efecto directo y primacía.

Además, el RIA es una norma con múltiples finalidades, entre ellas mejorar el funcionamiento del mercado interior de la UE a través de normas armonizadas para la introducción en el mercado, la puesta en servicio y la utilización de sistemas de IA en la UE, aunque también garantizar un elevado nivel de protección de la salud, la seguridad y los derechos fundamentales, incluidos la democracia, el estado de derecho y la protección del medio ambiente,

10 En realidad, se trata de una norma pertinente a efectos del Espacio Económico Europeo, lo que quiere decir que también se aplica en Noruega, Islandia y Liechtenstein.

y prestar apoyo a la innovación. Sin embargo, el Consejo de Europa sólo debía centrarse en garantizar que los sistemas de IA fueran plenamente compatibles con los derechos humanos, la democracia y el estado de derecho.

VALCKE y HENDRICKX argumentaron que el valor añadido del Convenio debía ser que, mientras que el RIA se centra en el mercado único digital y no crea nuevos derechos para las personas, el Convenio podría llenar estos vacíos al ser el primer tratado jurídicamente vinculante sobre IA que se centra en la democracia, los derechos humanos y el estado de derecho y, además, llegar a más Estados. Las autoras defendían que las dos normas no chocarían, sino que se complementarían porque el Convenio podría abordar los elementos que faltan en el RIA.[11] La duda es si esto ha sido así, viendo el resultado final de la negociación.

COTINO explica que el Convenio introduce la lírica de la protección de los derechos y los principios en la prosa del RIA más técnico y detallado y alaba el potencial normativo e interpretativo del convenio junto con su valor simbólico y metajurídico.

ALMADA y RADU argumentan que el RIA tendrá un efecto Bruselas[12] secundario que socavará la ambición de la UE de extender sus valores en la gobernanza de la IA a nivel global, porque el RIA sigue la legislación de la UE de seguridad de los productos y sus disposiciones ofrecen una protección limitada a algunos de los valores que la política de la UE pretende proteger, como los derechos fundamentales. En relación con la negociación del Convenio, consideran que la UE ha decidió resolver los posibles conflictos con el enfoque del RIA empujando al Consejo de Europa hacia converger con él.[13]

HICKOK *et al.* también han criticado que la UE decidiera dar prioridad al RIA y la ambición de la Comisión Europea de internacionalizar el RIA a través del Convenio, forzando el proceso.[14] Criticaron que la Comisión pidiera a sus delegados que ralentizaran el trabajo sobre el Convenio hasta que se completara el RIA.[15] Además, en su opinión,

11 HENDRICKX, Victoria; VALCKE, Peggy. "The Council of Europe's road towards an AI Convention: taking stock". *KU Leuven, CiTIP Blog*, 09/02/2023. https://www.law.kuleuven.be/citip/blog/the-council-of-europes-road-towards-an-ai-convention-taking-stock/ (última consulta: 17/06/2024).

12 El término "efecto Bruselas" se utiliza para referirse a un fenómeno por el cual los estándares europeos terminan aplicándose en todo el mundo, porque las empresas quieren acceder al mercado único y terminan así aplicando y exportando sus normas a nivel global, siendo uno de los casos comúnmente usados como ejemplo el de la protección de datos personales. Véase BRADFORD, Anu, *The Brussels Effect. How the European Union rules the world*, Oxford University Press, 2020.

13 ALMADA, Marco; RADU, Anca. "The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy", *German Law Journal*, 2023, pp. 1-18. Véase Adenda de la Decisión del Consejo por la que se autoriza la apertura de negociaciones en nombre de la Unión Europea con vistas a un convenio del Consejo de Europa sobre inteligencia artificial, derechos humanos, democracia y Estado de Derecho, Doc. 14173/22, ADD 1.

14 HICKOK, Merve; ROTENBERG, Marc; CAUNES Karine, "The Council of Europe Creates a Black Box for AI Policy". *Verfassungsblog: On Matters Constitutional*, 24/01/2023. Disponible en: https://verfassungsblog.de/coe-black-box-ai/ (última consulta: 17/06/2024).

15 BERTUZZI, Luca, "EU Commission postponed AI treaty negotiations with further delays in sight", *Euractiv*, 05/10/2022. Disponible en: https://www.euractiv.com/section/digital/news/eu-commission-postponed-ai-treaty-negotiations-with-further-delays-in-sight/ (última consulta: 17/06/2024). Esto no es algo nuevo,

también fue problemática esta convergencia porque significó que muchos de los negociadores provinieran de los ministerios de economía y no de justicia, que son los expertos en derechos humanos.

4. CRÍTICAS RECIBIDAS POR EL CONVENIO

Si bien el Convenio ha sido bien recibido, no está exento de crítica y el balance es agridulce. VAN KOLFSCHOOTEN y SHACHAR han subrayado la importancia del Convenio y alabado que sea una regulación global para una tecnología que cruza fácilmente jurisdicciones, el enfoque basado en los derechos humanos con evaluación de impacto y la aproximación de ciclo de vida completo. No obstante, también han señalado desafíos como: la aproximación sectorial neutral, la falta de reflexión sobre los nuevos derechos humanos, ciertas cuestiones de definición, y el proceso de negociación.[16]

ROBERTS *et al.* han definido el Convenio como más prometedor que iniciativas voluntarias de *soft-law,* como las del G20 o la UNESCO, por ser jurídicamente vinculante, ya que se espera que los Estados que lo ratifiquen lo traduzcan a su legislación nacional. Sin embargo, también avisan de que la ratificación de las convenciones del Consejo de Europa ha sido históricamente lenta, lo que resulta problemático dado el rápido ritmo del desarrollo de la IA y que los Estados que no participan en la redacción de las convenciones de la organización también se han negado anteriormente a ratificarlas debido a una percepción de falta de legitimidad, y en lugar de ello han presionado por un proceso más representativo en la ONU.[17] Esto último parece poco deseable viendo las diferencias en lo que respecta a protección de los derechos humanos en las negociaciones en la ONU para un convenio contra el cibercrimen.[18]

A continuación, se exponen más en detalle algunas de las críticas recibidas.

---

dado que la UE ya hizo algo similar durante las negociaciones de la modernización del Convenio de protección de datos personales del Consejo de Europa. Véase GASCÓN MARCÉN, Ana, "La Unión Europea y los convenios internacionales elaborados en el marco del Consejo de Europa", en: *Interacciones entre el Derecho de la Unión Europea y el Derecho internacional público,* Tirant lo Blanch, 2023, pp. 227-242.

16 VAN KOLFSCHOOTEN, Hannah; SHACHAR, Carmel. "The Council of Europe's AI Convention (2023-2024): Promises and pitfalls for health protection", *Health Policy*, vol. 138, 2023.

17 ROBERTS, Huw, *et al*. "Global AI governance: barriers and pathways forward". *International Affairs*, 2024, vol. 100, no 3, pp. 1275-1286.

18 PETIT DE GABRIEL, Eulalia W., "Libertad de expresión y delitos de opinión a la luz de la futura convención internacional integral sobre la lucha contra la utilización de las tecnologías de la información y las comunicaciones con fines delictivos.", en: *La libertad de expresión asediada: Delitos de odio, delitos de opinión, censuras de Gobiernos y de empresas*. Editorial Aranzadi, 2023. pp. 405-453 y GASCÓN MARCÉN, Ana, "The Budapest Convention and the UN Cybercrime Convention negotiations", en: *Global Cybersecurity and International Law*. Routledge, 2024. pp. 174-192.

4.1. Proceso de negociación

La negociación fue más larga de lo esperado, entre otras razones, porque la UE buscó retrasarla para cuadrar mejor con su proceso interno de adopción del RIA. No obstante, la mayor crítica al proceso fue la falta de transparencia. Mientras el CAHAI había contado con las contribuciones de la sociedad civil y de la academia que habían enriquecido sus resultados, el CAI decidió incluirlos en sus sesiones plenarias, pero no en el grupo de trabajo que redactó el convenio en el que sólo participaron los Estados. Parece que ésta fue una condición impuesta por Estados Unidos.[19] Es llamativo si tenemos en cuenta que el Consejo de Europa ha defendido el principio *multi-stakeholder* en regulación de temas digitales y que en lo referido a gobernanza de la IA se pone mucho énfasis en la necesidad de transparencia e inclusión.

Es usual que sean los representantes de los Estados los que negocien los textos de los tratados internacionales, pero hay una tendencia en los últimos años a abrir estas negociaciones.

Hickok *et al.* han criticado que el Consejo de Europa haya creado el mismo una caja negra para la política de la IA[20], imagen que se suele usar para criticar la opacidad de los propios algoritmos que se trata de regular,[21] saltándose sus propias normas y principios.

4.2. Contenido general

El Supervisor Europeo de Protección de Datos mostro su preocupación por el altísimo nivel de generalidad de las disposiciones jurídicas del Convenio, denunciando que su naturaleza en gran medida declarativa conduciría a una aplicación divergente del mismo, socavando así la seguridad jurídica y su valor añadido.[22] Esto en parte se debe a que los enfoques muy alejados de los diferentes negociadores han llevado a que el texto final sea poco ambicioso al terminar prevaleciendo en muchos casos el mínimo común denominador.

19 BERTUZZI, Luca, "US obtains exclusion of NGOs from drafting AI treaty", *Euractiv*, 17/01/2023. Disponible en: https://www.euractiv.com/section/digital/news/us-obtains-exclusion-of-ngos-from-drafting-ai-treaty/ (última consulta: 17/06/2024).

20 *Op. cit.* Hickok *et al.*

21 PASQUALE, Frank, *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

22 SUPERVISOR EUROPEO DE PROTECCIÓN DE DATOS, *Statement in view of the 10th and last Plenary Meeting of the Committee on Artificial Intelligence (CAI) of the Council of Europe drafting the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Disponible en: https://www.edps.europa.eu/system/files/2024-03/EDPS-2024-06-Statement-on-the-draft-convention-on-AI_EN.pdf?trk=public_post_comment-text (última consulta: 17/06/2024).

4.3. Ámbito de aplicación

Una de las principales controversias durante la negociación fue el ámbito de aplicación del Convenio y si debía aplicarse sólo al sector público o también al privado. Básicamente, mientras que la Comisión Europea abogaba por obligaciones también para el sector privado,[23] Estados Unidos (apoyado por Israel, Japón, Reino Unido y Canadá) era partidario de dejar al sector privado fuera.[24]

Esto se puede ver como un intento de continuar con la tradición liberal estadounidense de una escasa regulación de las cuestiones digitales muy protectora de sus propias empresas[25] y también de alinear el Convenio con su propia legislación, ya que la *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* de 2023 crea básicamente obligaciones para las agencias estatales.

Finalmente, el Convenio se aplica a las actividades dentro del ciclo de vida de los sistemas de IA realizadas por autoridades públicas (o actores privados que actúen en su nombre) y cada Parte abordará los riesgos e impactos que surjan de las actividades por parte del resto de actores privados de manera conforme con el objeto y propósito del Convenio. Cada Parte especificará en una declaración al ratificar el Convenio cómo se propone implementar esta obligación, ya sea aplicando los principios y obligaciones establecidas en el Convenio a actividades de actores privados o adoptando otras medidas apropiadas.

Esto hace que se pierda gran parte de la utilidad del Convenio al crear un régimen fragmentado y mucho más débil para el sector privado. Se trata de un compromiso/cesión porque al ser muchas de las grandes compañías del sector estadounidenses era una prioridad no desincentivar la ratificación por parte de Estados Unidos (y el resto de Estados que le apoyaban). Además, esta opción se aleja de la propuesta por el CAHAI que recomendó que el objetivo del instrumento fuera garantizar la plena coherencia con el respeto de los derechos humanos, el funcionamiento de la democracia y el estado de derecho en el desarrollo, diseño y aplicación de sistemas de IA, independientemente de si estas actividades eran realizadas por actores públicos o privados. Es decir, se ha ido a un resultado mucho menos ambicioso.

23 Si bien esta opinión no era unánime dentro de la UE porque, en el marco del Consejo, Alemania, Francia, España, Chequia, Estonia, Irlanda, Hungría y Rumania pidieron a la Comisión cierta flexibilidad para llegar a un acuerdo global, por lo que conseguir más signatarios debería ser una prioridad en lugar de un convenio amplio con un apoyo internacional más limitado.

24 Las negociaciones fueron más complejas con propuestas de mecanismos de *opt-in* y *opt-out*. Véase BERTUZZI, Luca, "EU Commission's last-minute attempt to keep private companies in world's first AI treaty", 24/01/2024, *Euractiv.* Disponible en: https://www.euractiv.com/section/artificial-intelligence/news/eu-commissions-last-minute-attempt-to-keep-private-companies-in-worlds-first-ai-treaty/ (última consulta: 17/06/2024).

25 Véase BRADFORD, Anu. *Digital empires: The global battle to regulate technology.* Oxford University Press, 2023. La autora describe el modelo americano impulsado por el mercado, frente al europeo centrado en los derechos y el chino controlado por el Estado, si bien apunta a la decreciente influencia global del tecnolibertarismo estadounidense y a la globalización de los derechos digitales europeos a través del poder regulador.

Otro elemento que ha rebajado la utilidad del Convenio al limitar su ámbito de aplicación es que queda fuera lo relativo a la defensa nacional. Esto no es sorprendente porque el Consejo de Europa no tiene competencia en este ámbito, pero se ha ido más allá porque no se exigirá a las Partes que apliquen este Convenio tampoco a actividades relacionadas con la protección de sus intereses de seguridad nacional, en el entendimiento de que dichas actividades se realicen de manera compatible con el derecho internacional aplicable, incluidos los derechos humanos, y con respeto a sus instituciones y procesos democráticos. Es decir, se ha creado una excepción similar a la del RIA sin ser necesario porque la UE no cuenta con competencias en ese ámbito, pero la misma limitación no existe respecto del Consejo de Europa.

4.4. Críticas de la Asamblea Parlamentaria

La propia Asamblea Parlamentaria del Consejo de Europa (un órgano formado por parlamentarios de los Estados miembros) acogió calurosamente la finalización del proyecto de Convenio, pero lamentó que no cubriera en igual medida a los actores públicos y privados. Por ello pidió enérgicamente a todos los Estados miembros de la organización que, al ratificar el Convenio, optaran por aplicar plenamente sus disposiciones a las actividades de los actores privados. La Asamblea propuso varias enmiendas al proyecto de Convenio, como que fueran los Estados los que pudieran optar por no aplicarlo en cuestiones de seguridad o defensa nacional, siempre que se cumpliera con los derechos humanos; que los Estados deberían imponer limitaciones, o incluso prohibiciones, a ciertos usos de la IA considerados incompatibles con los derechos humanos (como en el RIA); y propusieron insertar una disposición específica sobre salud y medio ambiente.[26] No obstante, el texto no se enmendó para atender a sus propuestas.

5. CONCLUSIONES

El balance de este Convenio es ambivalente, por un lado, es un histórico paso adelante, porque es necesario crear obligaciones jurídicamente vinculantes para el desarrollo y la puesta en funcionamiento de sistemas de IA desde una perspectiva de protección de los derechos humanos, la democracia y el estado de derecho, yendo más allá del *soft-law* y la autorregulación con un acuerdo de un número elevado de países comprometidos con estos valores. No obstante, también se trata de una oportunidad perdida para haber llegado a un acuerdo más ambicioso y significativo, habiendo sacrificado excesivamente el contenido material y ámbito de aplicación por un mayor consenso.

26 ASAMBLEA PARLAMENTARIA DEL CONSEJO DE EUROPA, *Opinion 303 (2024) on the Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Disponible en: https://pace.coe.int/en/files/33517/html (última consulta: 17/06/2024).

Ahora queda esperar para ver cuántos países efectivamente ratifican el Convenio y si la Conferencia de las Partes tiene realmente un papel proactivo impulsando los avances en la materia para poder analizar su impacto real.

6. BIBLIOGRAFÍA

ALMADA, Marco; RADU, Anca. "The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy", *German Law Journal*, 2023, p. 1-18.

ASAMBLEA PARLAMENTARIA DEL CONSEJO DE EUROPA, *Opinion 303 (2024) on the Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Disponible en: https://pace.coe.int/en/files/33517/html (última consulta: 17/06/2024).

BERTUZZI, Luca, "EU Commission postponed AI treaty negotiations with further delays in sight", *Euractiv*, 05/10/2022. Disponible en: https://www.euractiv.com/section/digital/news/eu-commission-postponed-ai-treaty-negotiations-with-further-delays-in-sight/ (última consulta: 17/06/2024).

BERTUZZI, Luca, "EU Commission's last-minute attempt to keep private companies in world's first AI treaty", 24/01/2024, *Euractiv.* Disponible en: https://www.euractiv.com/section/artificial-intelligence/news/eu-commissions-last-minute-attempt-to-keep-private-companies-in-worlds-first-ai-treaty/ (última consulta: 17/06/2024).

BERTUZZI, Luca, "US obtains exclusion of NGOs from drafting AI treaty", *Euractiv*, 17/01/2023. Disponible en: https://www.euractiv.com/section/digital/news/us-obtains-exclusion-of-ngos-from-drafting-ai-treaty/ (última consulta: 17/06/2024).

BRADFORD, Anu. *Digital empires: The global battle to regulate technology.* Oxford University Press, 2023.

BRADFORD, Anu, *The Brussels Effect. How the European Union rules the world*. Oxford University Press, 2020.

CAHAI, *Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law*, CAHAI(2021)09rev. Disponible en: https://rm.coe.int/cahai-2021-09rev-elements/1680a6d90d (última consulta: 17/06/2024).

COMMITTEE OF EXPERTS ON INTERNET INTERMEDIARIES (MSI-NET), *Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications,* DGI(2017)12;

COTINO HUESO, Lorenzo, "The Council of Europe's Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law", *Rivista Interdisciplinare Sul Diritto Delle Amministrazioni Pubbliche*. n. 2/2024. Disponible en: https://ceridap.eu/the-council-of-europes-convention-on-artificial-intelligence-human-rights-democracy-and-the-rule-of-law/?lng=en (última consulta: 23/09/2024).

GASCÓN MARCÉN, Ana, "Derechos humanos e inteligencia artificial", en: *Setenta años de Constitución Italiana y cuarenta años de Constitución Española*, Vol. 5, BOE y Centro de Estudios Políticos y Constitucionales, 2020, pp. 335-350.

GASCÓN MARCÉN, Ana, "The Budapest Convention and the UN Cybercrime Convention negotiations", en: *Global Cybersecurity and International Law*. Routledge, 2024. pp. 174-192.

GASCÓN MARCÉN, Ana, "La Unión Europea y los convenios internacionales elaborados en el marco del Consejo de Europa", en: *Interacciones entre el Derecho de la Unión Europea y el Derecho internacional público*, Tirant lo Blanch, 2023, pp. 227-242.

GSTREIN, Oskar Josef, "The Council of Europe as an Actor in the Digital Age: Past Achievements", *Future Perspectives. Festschrift der Mitarbeiter* Innen und Doktorand* Innen zum*, 60, 2019, pp. 77-90.

HENDRICKX, Victoria; VALCKE, Peggy. "The Council of Europe's road towards an AI Convention: taking stock". *KU Leuven, CiTIP Blog*, 09/02/2023. https://www.law.kuleuven.be/citip/blog/the-council-of-europes-road-towards-an-ai-convention-taking-stock/ (última consulta: 17/06/2024).

HICKOK, Merve; ROTENBERG, Marc; CAUNES Karine, "The Council of Europe Creates a Black Box for AI Policy". *Verfassungsblog: On Matters Constitutional*, 24/01/2023. Disponible en: https://verfassungsblog.de/coe-black-box-ai/ (última consulta: 17/06/2024).

LESLIE, David; BURR, Christopher; AITKEN, Mhairi; COWLS, Josh; KATELL, Mike; BRIGGS, Morgan, *Artificial intelligence, human rights, democracy, and the rule of law: a primer*. Council of Europe and The Alan Turing Institute, 2021.

MANTELERO, Alessandro; FANUCCI, Francesca, "Great Ambitions. The International Debate on AI Regulation and Human Rights in the Prism of the Council of Europe's CAHAI", *European Yearbook on Human Rights* 2022. Cambridge, Intersentia, pp. 225-252.

PASQUALE, Frank, *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

PETIT DE GABRIEL, Eulalia W., "Libertad de expresión y delitos de opinión a la luz de la futura convención internacional integral sobre la lucha contra la utilización de las tecnologías de la información y las comunicaciones con fines delictivos.", en: *La libertad de expresión asediada: Delitos de odio, delitos de opinión, censuras de Gobiernos y de empresas*. Editorial Aranzadi, 2023. pp. 405-453

ROBERTS, Huw; HINE, Emmie; TADDEO, Mariarosaria; FLORIDI, Luciano "Global AI governance: barriers and pathways forward". *International Affairs*, 2024, vol. 100, n. 3, pp. 1275-1286.

SUPERVISOR EUROPEO DE PROTECCIÓN DE DATOS, *Statement in view of the 10th and last Plenary Meeting of the Committee on Artificial Intelligence (CAI) of the Council of Europe drafting the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Disponible en: https://www.edps.europa.eu/system/files/2024-03/EDPS-2024-06-Statement-on-the-draft-convention-on-AI_EN.pdf?trk=public_post_comment-text (última consulta: 17/06/2024).

VAN KOLFSCHOOTEN, Hannah; SHACHAR, Carmel. "The Council of Europe's AI Convention (2023-2024): Promises and pitfalls for health protection", *Health Policy*, vol. 138, 2023.

Capítulo 12

CONSIDERACIONES CLIMÁTICAS Y EL REGLAMENTO DE INTELIGENCIA ARTIFICIAL.

Elena Cisneros Cabrerizo
Personal investigador en formación[1]
Universidad de Zaragoza

RESUMEN: Esta comunicación analiza las consideraciones ambientales que aparecen en el Reglamento Europeo de Inteligencia Artificial, reflexionando sobre el proceso de negociaciones que llevó a su aprobación. En estas discusiones se desarrollaron cuestiones ambientales (tales como la transparencia o el coste energético de la IA) que no se han visto incluidas en la redacción final del reglamento. La mitigación de la relevancia medioambiental en el texto final plantea una serie de consecuencias y limitaciones en el ámbito de la transición digital y ecológica. A lo largo del texto se estudia el equilibrio de intereses medioambientales en el campo de la IA como una problemática que debe valorarse desde la perspectiva jurídica, ya que estos sistemas se presentan como herramientas útiles de gestión y planificación en este ámbito y, simultáneamente, como tecnologías de gran impacto ambiental.

PALABRAS CLAVE: Inteligencia artificial – Impacto ambiental – Transición digital – Transición ecológica – Unión Europea – Reglamento de Inteligencia Artificial

1 Investigadora predoctoral de la Universidad de Zaragoza, en el Área de Derecho Internacional Público y Relaciones Internacionales, gracias a la financiación y el apoyo de la Fundación Ramón Areces.

1. INTRODUCCIÓN

La relación entre el cambio climático y la inteligencia artificial tiene una doble naturaleza: por un lado, la IA es -o puede ser- una herramienta mitigadora de los efectos del cambio climático; por otro lado, las ramificaciones ambientales de su aplicación son una cuestión a la que es necesario hacer referencia. Independientemente del enfoque en el que elijamos centrarnos, la conexión entre ambas cuestiones es una realidad inevitable a tratar al acercarse a esta incipiente tecnología, especialmente en la dimensión europea, debido las políticas de transición ecológica y digital de la Unión Europea.

Durante la última legislatura (2019-2024), la Comisión Europea empezó a desarrollar una fuerte estrategia digital buscando el fortalecimiento de la autonomía de la Unión Europea en este ámbito, ligada a una concepción de transiciones gemelas entre la transición digital y la ecológica[2]. La decisión de acercarse a la inteligencia artificial desde los postulados del Pacto Verde ha sido una seña de identidad en la estrategia europea en la materia y la relación entre ambas políticas se ha presentado como una de las principales prioridades de la última legislatura. En el discurso sobre el Estado de la Unión de 2020 se presentaron algunas de las líneas maestras para enfocar el desarrollo legislativo en la próxima década y afrontar los nuevos retos planteados en el panorama social; en este marco se expresó la voluntad de situar el ámbito digital en el centro del desarrollo de la Unión[3]. Dentro de las cuatro principales líneas de cara a futuro se remarcó la importancia de desarrollar en particular el sector europeo de la inteligencia artificial, mencionado la propuesta de una ley específica en la cuestión que ha conducido a la aprobación del Reglamento de Inteligencia Artificial.

La asociación entre las políticas ambientales y digitales se ha ido desarrollando a partir del paralelismo que se ha trazado entre la transición digital y transición ecológica, señalándolas como ámbitos vertebradores de la política legislativa en el marco 2030. El objetivo de priorizar el desarrollo digital se ha manifestado en la política relativa a *la década digital* en la que la Unión Europea ha ido creando una densa infraestructura normativa. Las medidas legislativas tomadas en este contexto -el Reglamento de Gobernanza de Datos[4], Reglamento de Servicios Digitales[5] o Reglamento de Mercados Digitales[6]- han situado de manera paulatina a

2 *Shaping Europe's digital future*. (s. f.). European Commission. https://commission.europa.eu/strategy- and-policy/priorities-2019-2024/europe-fit-digital-age/shaping-europes-digital-future_en

3 *Estado de la Unión 2020*. 16 de septiembre de 2020. https://state-of-the-union.ec.europa.eu/state-union-2020_es

4 Reglamento(UE) 2022/868 del Parlamento Europeo y del Consejo de 30 de mayo de 2022 relativo a la gobernanza europea de datos y por el que se modifica el Reglamento (UE) 2018/1724 (Reglamento de Gobernanza de Datos) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R0868

5 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) https://eur-lex.europa.eu/legal-content/EN/TXT/?toc=OJ%3AL%3A2022%3A277%3ATOC&uri=uriserv%3AOJ.L_.2022.277.01.0001.01.ENG

6 Reglamento (UE) 2022/1925 del Parlamento Europeo y del Consejo de 14 de septiembre de 2022 sobre mercados disputables y equitativos en el sector digital y por el que se modifican las Directivas (UE) 2019/1937 y (UE) 2020/1828 (Reglamento de Mercados Digitales) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1925

la Unión Europea en la vanguardia de la creación de una estrategia digital con una marcada identidad propia que pretende no solo definir el futuro de la Unión sino tener un efecto más allá de sus fronteras[7].

En la definición de esta estrategia digital común, el avance de la transición ecológica y su estrecha relación con el proceso digital es una decisión intencionada y un motivo recurrente en ambas políticas. En el *Informe de la prospectiva estratégica 2022: Hermanamiento de las transiciones ecológica y digital en el nuevo contexto geopolítico*[8] se delimita la relación entre ellas, entendiendo el papel clave de las nuevas tecnologías como herramientas para mitigar los efectos del cambio climático, pero también señalando el peligro del aumento del uso de tecnologías digitales cuya consecuencia será el incremento del consumo de energía y la generación de un mayor volumen de residuos, así como el aumento de la huella medioambiental digital por el crecimiento de su uso.

Este marco estratégico señala la voluntad de la Unión no solo de mantener un papel de liderazgo en la regulación de la esfera digital, sino de dotar este liderazgo de una nota particular, incluyendo la vocación medioambiental en su regulación. La ambición por reafirmar esta identidad propia se aprecia en el desarrollo de la estrategia de la IA de la UE, que se encuadra en este contexto y aparece como un elemento más a considerar dentro de este entramado ecológico y digital. Uno de los primeros documentos de esta estrategia, *Directrices Éticas para una IA fiable*[9], presenta la necesidad del desarrollo de un concepto de IA en el que la licitud, la ética y la robustez se posicionen en el núcleo de todos los sistemas, objetivos que van a definir la visión europea de la IA. En su enumeración de principios las consideraciones medioambientales se encuentran incluidas desde la perspectiva de la sostenibilidad y el respeto al medio ambiente. Este primer acercamiento de la UE al concepto de *IA sostenible* se concreta en cuestiones requerimientos de seguridad general para los sistemas, entendiendo que un sistema no es seguro en caso de que pueda causar daños medioambientales; garantías y mecanismos de evaluación en su desarrollo y uso, que existan controles tanto en la cadena de suministros como en su funcionamiento respecto al impacto ambiental.

Las conclusiones son el fruto de los primeros pasos que intentan definir las notas de la IA para dotarla de la perspectiva europea que se caracteriza por señalar la necesidad de crear un perfil de sistemas de IA más humanos, en el que la ética sea un elemento fundamental y cuyo funcionamiento esté claramente alineado con los objetivos preexistentes de la Unión. Dentro de ese concepto de ética, se recogen una serie de elementos diversos que se van desarrollando desde las primeras comunicaciones del año 2020 hasta la actualidad que siguen trazando el paralelismo entre las dos transiciones -haciendo hincapié en la necesidad de que la Unión Europea establezca este novedoso marco regulatorio- y la necesidad de prestar especial atención

7 Report 2030 Digital Decade, Report in the State of the State of the Digital Decade 2024 Communication, European Commission https://digital-strategy.ec.europa.eu/en/policies/2024-state-digital-decade-package

8 Comunicación de la Comisión al Parlamento Europeo y al Consejo Informe de prospectiva estratégica 2022 Hermanamiento de las transiciones ecológica y digital en el nuevo contexto geopolítico https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52022DC0289

9 Comisión Europea, Dirección General de Redes de Comunicación, Contenido y Tecnologías, Directrices éticas para una IA fiable, Oficina de Publicaciones, 2019, https://data.europa.eu/doi/10.2759/14078

a conceptos como la sostenibilidad, el respeto por el medio ambiente o la transición ecológica justa. Debido a la naturaleza de las directrices, que sirvieron como un ejercicio de estudio preparatorio en el que se planteaba el futuro marco regulatorio de la IA; y a su contenido, que señala que las consideraciones relativas al medio ambiente son indisolubles del concepto de ética, era una expectativa razonable que el posterior desarrollo normativo del RIA reflejase estas consideraciones, quizás no de manera directa, pero sí considerable y que situase la fiabilidad como un motivo distintivo de carácter europeo necesario en los sistemas de inteligencia artificial para garantizar que su desarrollo y su empleo sean éticos.

2. REGLAMENTO DE INTELIGENCIA ARTIFICIAL

El 12 de julio de 2024 se publicó oficialmente el texto del Reglamento de Inteligencia Artificial (RIA) en el DOUE[10] y, pese al valor que se había dado a la responsabilidad medioambiental y a la creación de un marco de protección del medio ambiente, estas consideraciones han quedado muy reducidas, lo que abre un escenario incierto de cara a la lucha contra el cambio climático y al desarrollo tecnológico de la UE.

En el texto final la preocupación por el medio ambiente se refleja de manera muy limitada, tanto por su localización dentro del texto, como por la profundidad de los compromisos que se han alcanzado. En las consideraciones iniciales aparecen mencionadas de manera regular, equiparando su importancia con la de la protección de los derechos fundamentales o de la democracia. Estas tres ideas se presentan de manera conjunta en numerosas ocasiones en el preámbulo, posicionándolas en un nivel de atención equiparable. Sin embargo, este interés se ve considerablemente diluido a lo largo del articulado, en el que las referencias al medio ambiente y a la protección medioambiental se reducen, y en aquellos momentos en los que se mencionan se encuentran separadas de las consideraciones relativas a las evaluaciones de riesgos. La exclusión del impacto ambiental de la clasificación de riesgos limita enormemente la eficacia de las previsiones ambientales que sí que se recogen en el texto, ya que éstas no se encuentran enmarcadas en un régimen de control o seguimiento firme que cree obligaciones para el desarrollo o puesta en marcha de los sistemas. Esta decisión implica la falta una protección diferenciada del medio ambiente y excluye las consideraciones climáticas de los mecanismos de evaluación de derechos fundamentales. La disonancia entre la aparición constante de la cuestión ambiental en el preámbulo y el tratamiento muy limitado que obtiene en el articulado refleja un trato irregular ya que no se ha establecido un sistema de protección medioambiental, pero se ha incluido el concepto de impacto o protección medioambiental de manera diferenciada junto con otros objetivos prioritarios de la Unión en el preámbulo.

10 Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) nº 300/2008, (UE) nº 167/2013, (UE) nº 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-2024-81079

Las limitaciones que presenta el articulado del RIA no implican que las consideraciones climáticas desaparezcan por completo: un ejemplo de las previsiones que se contemplan se encuentra en el artículo 46 en el que se reconoce la posibilidad de introducir sistemas de alto riesgo en el mercado -de manera excepcional- con el fin de proteger una serie de valores fundamentales, entre ellos el medio ambiente, lo que mantiene el espíritu de tratamiento diferenciado que aparecía en el preámbulo. Por otra parte, el artículo 59 del Reglamento reconoce que -también de forma excepcional y en un espacio controlado de pruebas- se podrá proceder a tratar datos personales con la finalidad de entrenar un sistema cuando el sistema que se esté desarrollando se refiera a algún ámbito que merece un especial apoyo, entre ellos, la protección y mejora de la calidad del medio ambiente, así como la sostenibilidad energética.

Estos artículos son dos ejemplos claros de que la Unión Europea mantiene este interés en el desarrollo de una infraestructura sólida de sistemas de inteligencia artificial destinados a la gestión y protección medioambiental. Es más, de las notas de estos artículos se aprecia como la protección medioambiental mantiene una consideración diferenciada como ámbito de valor fundamental, ya que se establecen excepciones para poder emplear los sistemas con finalidades ambientales, pero que a la hora de la creación de un marco de protección sólido su importancia ha mermado significativamente frente a las propuestas que se hicieron a lo largo del proceso de negociación del RIA.

2.1. Propuesta del Parlamento Europeo

La forma en la que el texto final del reglamento se refiere a las cuestiones medioambientales evidencia un choque entre dos visiones dentro de la Unión Europea: la primera visión -reflejada en los documentos preparatorios y en la propuesta del Parlamento Europeo- en la que el impacto medioambiental es una preocupación constante a lo largo del articulado, frente a una segunda visión contenida en la propuesta de la Comisión, que reduce esta importancia y aparece en la versión final.

La evidencia de la enrome brecha entre estas dos concepciones de cómo debe regularse la inteligencia artificial puede verse en la comparación del texto final con la propuesta que hizo el Parlamento Europeo. La variación en la consideración del medio ambiente, así como en el número y efectividad de las medidas para combatir el impacto ambiental de la IA se ve a lo largo de varios puntos de estos textos, siendo relevante como introducción el contenido del artículo 4.a) de la propuesta del Parlamento Europeo que establecía los principios generales aplicables a todos los sistemas de IA, reconociendo en su apartado f) que los sistemas deben desarrollarse y usarse de manera sostenible y respetuosa con el medio ambiente. La preocupación del impacto medioambiental de los sistemas de alto riesgo se refleja también en el propuesto artículo 12.2.a) en el que se indica que:

> Los sistemas de IA de alto riesgo deben diseñarse y desarrollarse con capacidades de registro que permitan la grabación del consumo de energía, la medición o el cálculo del uso de recursos y el impacto ambiental… durante todas las fases del ciclo de vida del sistema.

En esta misma línea el artículo 29.a) propuesto por el Parlamento hacía referencia a la evaluación de impacto en derechos fundamentales para los sistemas de alto riesgo entre cuyos requisitos mínimos se exigía de acuerdo con su apartado g):

> el impacto adverso razonablemente previsible del uso del sistema en el medio ambiente;

La inclusión de esta valoración en la evaluación de impacto en derechos fundamentales es quizás una de las mayores pérdidas en el texto final, ya que no solo se trata de una herramienta efectiva para mitigar el impacto ambiental de los sistemas, sino que además ayuda a mantener esta coherencia interna al mantener en el articulado la relación entre las consideraciones de derechos fundamentales y la cuestión medioambiental. En esta misma línea y profundizando en el marco de protección al medio ambiente el artículo 13.3.b) iii) incluía también consideraciones medioambientales, señalando que el daño ambiental era un motivo suficiente para cesar el uso de un sistema de alto riesgo en caso de mal uso, el artículo 14.2 señalaba que la supervisión humana debe minimizar el riesgo de impacto en la salud, los derechos fundamentales o el medio ambiente y el artículo 28, establecía que los proveedores deberán demostrar que se han identificado y mitigado riesgos razonables para la salud, derechos fundamentales o medio ambiente.

Estos artículos manifiestan la reducción de la preocupación medioambiental en el texto de la versión final, en el que éstas han llegado recudidas y circunscritas al ámbito del desarrollo de códigos de conducta (artículo 95 RIA) lo que les resta efectividad al tratarse de mecanismos de cumplimiento voluntarios, o a la previsión de posteriores revisiones del texto que ofrezcan una visión más comprensiva una vez éste haya entrado en vigor (artículo 112 RIA). El artículo 112 sobre evaluación y revisión es quizás uno de los elementos más prometedores de este reglamento ya que, pese a algunas de las carencias ya señaladas, la posible revisión posterior ofrecería una herramienta para profundizar en la protección medioambiental.

2.2. Limitaciones de la versión final y falta de consideración medioambiental

Al comparar la propuesta del Parlamento con la versión final del texto, que va a definir la posición europea en este ámbito, nos encontramos con un marco de regulación ambiental imperfecto. El legislador europeo ha entendido que la IA es una herramienta de gestión y de planificación medioambiental, y las referencias en el articulado al medio ambiente se encaminan en la dirección de encauzar la IA como herramienta técnica para lograr estos fines, pero la visión parcial de su empleo excluye un elemento fundamental: el impacto de esta tecnología en el cambio climático. En el texto las referencias a herramientas de control de este impacto se recogen en el artículo 40, con la previsión de la armonización de estándares y en el Anexo XI en relación al artículo 53 respecto a la documentación técnica a presentar por los desarrolladores de sistemas de uso general incluye en su apartado 2d *los recursos computacionales utilizados para entrenar el modelo* y en el 2e hace referencia expresa al consumo de energía.

La decisión de no seguir la línea del PE para abordar esta cuestión es desafortunada, ya que la inteligencia artificial, independientemente de su uso, presenta impacto directo e

indirecto en el medio ambiente, tal y como expresa la OCDE en su informe de 2022[11]. Las distintas dimensiones de este impacto se manifiestan en todas las fases de su existencia: en la producción de los sistemas, en el transporte, en su vida activa y en el fin de su ciclo de vida. Las estimaciones de este informe expresan que, únicamente en su fase de funcionamiento, el consumo energético de los sistemas representa el 1% del consumo energético global, porcentaje que con el aumento de su uso se verá incrementado. De este dato, por preocupante que sea, no es posible desglosar el porcentaje de consumo que corresponde a la Unión Europea, ya que la opacidad y falta de datos públicos del consumo de recursos de los sistemas limitan los elementos de estudio. La imposibilidad de poder obtener esta información dificulta el estudio del sector a nivel europeo, así como la presentación de propuestas de mejora de la eficiencia. La posibilidad de establecer mecanismos para medir el consumo de recursos y llevar a cabo un seguimiento fiel del mismo a nivel europeo supondría una medida que, sin ser extraordinariamente invasiva, permitiría el seguimiento, la cooperatividad y fomentaría un mayor nivel de transparencia estableciendo un precedente positivo para el desarrollo del sector. Las referencias en el RIA en el Anexo IX respecto a las obligaciones de documentación podrían ser una medida útil en este sentido, pero se ven muy diluidas al no tratarse de obligaciones de presentación o depósito de los datos.

Ligada a esta cuestión cabe señalar el vigente artículo 27 del Reglamento de Inteligencia Artificial que hace referencia a la evaluación de impacto relativa a los derechos fundamentales para los sistemas de IA de alto riesgo. El texto final del artículo dista de la propuesta del PE en su artículo 29, ya que ésta recogía la importancia de que las consideraciones medioambientales formen parte de las cuestiones que deben ser evaluadas. La inclusión de este enfoque permitía tener herramientas para evaluar la trazabilidad del impacto que presenta la implantación de sistemas de IA y permitía la obtención de un mayor volumen de datos de comparación, favoreciendo la transparencia y la creación de un marco de estudio comparativo de la eficiencia medioambiental.

El ejemplo de la oportunidad climática perdida podemos encontrarlo en el Reglamento de Servicios Digitales, antecedente que resulta relevante para analizar este fenómeno, ya que su articulado ofrece una serie de herramientas que se podrían haber traspasado al RIA. En el RSD se pone el foco en los grandes prestadores de servicios, que pueden considerarse como gatekeepers, y sujetos definidores del mercado[12] a los cuales se les impone la obligación de mitigar aquellos riesgos que se consideren sistémicos en el funcionamiento de sus servicios[13]. En un marco que, al igual que el RIA, posiciona el riesgo en el centro de su normativa, el RSD ofrece un mecanismo en el cual se pueden incluir las consideraciones medioambientales, lo que podría haber sido una herramienta útil para implementar en el Reglamento de

11 OECD (2022), "Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint", *OECD Digital Economy Papers*, No. 341, OECD Publishing, Paris, https://doi.org/10.1787/7babf571-en.

12 Griffin, R. (2023). Climate Breakdown as a Systemic Risk in the Digital Services Act. opus4.kobv.de. https://doi.org/10.48462/opus4-5075

13 Artículo 34 y 35 Reglamento de Servicios Digitales https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32022R2065

Inteligencia Artificial sin tener que llegar a la rotundidad de los postulados propuestos por el Parlamento.[14]

2.2.1. Propuestas de terceros y perspectivas de implantación

La cuestión medioambiental en el campo de la IA representa un desafío al que diversos actores están tratando de acercarse: Global Partnership on AI Report, Climate Change AI y el Centre for AI and Climate emitieron un informe de recomendaciones sobre cómo tratar la IA desde la perspectiva medioambiental, como herramienta mitigadora pero también haciendo referencia al impacto de su uso.[15] La evaluación de impacto climático se presenta como una de las principales herramientas que, junto a la elaboración de informes, permiten entender el impacto que tienen estas tecnologías y desarrollar una política de adaptación que se adecúe a los resultados obtenidos. Los mayores retos para llevar a cabo estas estrategias de evaluación y estudio radican en la falta de información sobre el consumo de los sistemas existentes, por lo que es necesario contar con un marco de mayor transparencia a raíz del cual se puedan obtener estos datos, requisito necesario para garantizar el desarrollo sostenible del sector y que subraya las carencias del RIA que no ha establecido un marco de obligaciones más sólido en este ámbito.

Pese a todas estas propuestas es necesario señalar que tratar el impacto medioambiental de la IA requiere un enfoque supranacional, que permita estudiar la magnitud de su impacto y desarrollar las medidas de mitigación adecuadas[16] que no se limiten a un momento concreto de la vida de un sistema, sino que abarquen la totalidad de su funcionamiento, ya que la compartimentalización va en detrimento de su efectividad[17]. Siguiendo esta línea la OCDE indica la necesidad de entender que este impacto medioambiental abarca las diferentes fases de la vida de un sistema, desde la producción, el entrenamiento, la utilización y el fin de la vida de este e intentar favorecer el desarrollo de un estándar de sostenibilidad más alto en cada una de estas fases.

14 El impacto del RSD en el ámbito de la IA está todavía por determinar, ya que en el momento de incluir servicios de IA estos grandes prestadores van a tener que incluir una evaluación de impacto en las que estas cuestiones queden cubiertas.

15 Global Partnership on AI Report Climate Change and AI Recommendations for Government Action https://www.c-ai-c.org/climate-change-ai-report

16 GPAI 2023. Responsible AI Strategy for the Environment: Workshop Report, November 2023, Global Partnership on AI. https://gpai.ai/projects/responsible-ai/Responsible%20AI%20Strategy%20for%20the%20Environment%20-%20Workshop%20Report.pdf

17 Sanders, Barrie *At the Intersection of Climate Change, AI, and Human Rights Law: Towards a Solidarity-Based Approach (Part 1)*. (s. f.-b). Afronomicslaw.org, 17 November 2023 https://www.afronomicslaw.org/category/analysis/intersection-climate-change-ai-and-human-rights-law-towards-solidarity-based

3. CONCLUSIONES

El Reglamento de Inteligencia Artificial se presenta ahora como una normativa de enorme valor, al ser un texto comprensivo que ofrece un marco regulatorio centrado en la protección y en el desarrollo de unos sistemas éticos. Pese a ser un gran avance legislativo para la estrategia digital de la Unión Europea el desarrollo de la protección medioambiental ha quedado muy limitado en el texto y plantea dudas respecto a la efectividad que tendrá el reglamento para poder tratar algunos de los problemas intrínsecos al funcionamiento de la IA ya que no existen herramientas de evaluación de impacto ambiental, ni de mitigación y las obligaciones de transparencia en este ámbito han quedado considerablemente reducidas. El RIA, pese a su importancia como herramienta legislativa representa también una oportunidad perdida de la Unión para posicionar las consideraciones medioambientales en la vanguardia de su estrategia de IA.

La eficacia de los mecanismos jurídicos para mitigar el impacto de los sistemas en el medio ambiente es una cuestión por analizar, puesto que el desarrollo constante de la IA y la falta de precedentes dificultan esta evaluación, pero el acercamiento que hace el RIA de la cuestión medioambiental es deficiente. El reglamento no rechaza completamente la importancia de las cuestiones ambientales -la posibilidad de revisión del artículo 112 así como la previsión de un proceso de armonización a través de la estandarización, artículo 40- pero sí que opta por un acercamiento parcial y muy limitado, desconectado de las líneas discursivas de la UE en los últimos años. En el RIA vemos la falta de un tratamiento conjunto de la transición ecológica y la digital después pese a establecer la IA como una herramienta necesaria y fundamental para la realización del Pacto Verde. La falta de previsión del impacto en el cambio climático de la IA genera un desequilibrio entre esas dos "transiciones hermanadas" y abre un escenario en el que se rechaza hacer un seguimiento activo del impacto que tiene la IA en el cambio climático mientras se pretende que ésta sea una de las principales herramientas para alcanzar los objetivos climáticos de la Unión.

Conforme la popularización de los sistemas de IA siga creciendo el desarrollo y su consiguiente impacto medioambiental van a continuar incrementado, por lo que es necesario buscar vías para completar las limitaciones del reglamento y ofrecer un sistema de protección ambiental que analice los principales focos de preocupación -tanto dentro del funcionamiento de los sistemas como dentro de las diferentes formas de impacto ambiental- y permita desarrollar una estrategia ambiental y digital verdaderamente cohesionada que cumpla con el objetivo de desarrollar ambas transiciones de manera paralela.

4. BIBLIOGRAFÍA

AI and Sustainability: Will AI Help or Perpetuate the Climate Crisis? (2022, 19 Septiembre). Stanford HAI. https://hai.stanford.edu/news/ai-and-sustainability-will-ai-help-or-perpetuate-climate-crisis

AI's carbon footprint problem. (2020, 2 julio). Stanford HAI. https://hai.stanford.edu/news/ais-carbon-footprint-problem

Analyzing the European Union AI Act: What Works, What Needs Improvement. (2023, 21 julio). Stanford HAI. https://hai.stanford.edu/news/analyzing-european-union-ai-act-what-works-what-needs-improvement

Comisión Europea, Directrices éticas para una IA fiable, Oficina de Publicaciones, 2019, https://data.europa.eu/doi/10.2759/14078

COMUNICACIÓN DE LA COMISIÓN AL PARLAMENTO EUROPEO Y AL CONSEJO Informe de prospectiva estratégica 2022 Hermanamiento de las transiciones ecológica y digital en el nuevo contexto geopolítico https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52022DC0289

Crawford, Kate (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

Ecosystem Graphs: The Social Footprint of Foundation Models. (2023, 29 marzo). Stanford HAI. https://hai.stanford.edu/news/ecosystem-graphs-social-footprint-foundation-models

Estado de la Unión 2020. (s. f.). Estado de la Unión. https://state-of-the-union.ec.europa.eu/state-union-2020_es

Global Partnership on AI Report Climate Change and AI Recommendations for Government Action https://www.c-ai-c.org/climate-change-ai-report

GPAI 2023. Responsible AI Strategy for the Environment: Workshop Report, November 2023, Global Partnership on AI. https://gpai.ai/projects/responsible-ai/Responsible%20AI%20Strategy%20for%20the%20Environment%20-%20Workshop%20Report.pdf

Griffin, Rachel (2023). Climate Breakdown as a Systemic Risk in the Digital Services Act. opus4.kobv.de. https://doi.org/10.48462/opus4-5075

Hacker, Philipp, Sustainable AI Regulation (June 1, 2023). Disponible en SSRN: https://ssrn.com/abstract=4467684 or http://dx.doi.org/10.2139/ssrn.4467684

Industry Brief | Sustainability and AI. (s. f.). Stanford Institute For Human-Centered Artificial Intelligence. https://hai.stanford.edu/industry-brief-sustainability-and-ai

Li Pengfei, Yang Jianyi,. Islam Mohammad A & Ren Shaolei. (2023, 6 abril). *Making AI Less «Thirsty»: Uncovering and Addressing the Secret Water Footprint of AI Models*. arXiv.org. https://arxiv.org/abs/2304.03271

Novelli, Claudio, Casolari, Federico, Rotolo, Antonino, Taddeo, Mariarosaria, & Floridi, Luciano (2023). Taking AI risks seriously: a new assessment model for the AI Act. *AI & Society*. https://doi.org/10.1007/s00146-023-01723-z

Pagallo, Ugo, Ciani Sciolla, Jacopo and Durante, Massimo (2022), «The environmental challenges of AI in EU law: lessons learned from the Artificial Intelligence Act (AIA) with its drawbacks», *Transforming Government: People, Process and Policy*, Vol. 16 No. 3, pp. 359-376. https://doi.org/10.1108/TG-07-2021-0121

Rolnick, David, Donti, Priya L, Kaack, Lynn. H., Kochanski, Kelly., Lacoste, Alexandre., Sankaran, Kris., Ross, Anderew. S., Milojevic-Dupont, Nikola, Jaques, Natasha, Waldman-Brown, Anna, Luccioni, Alexandra., Maharaj, Tegan, Sherwin, Evan D., Mukkavilli, S. Karthik, Kording, Konrad. P., Gomes, Carla, Ng, A. Y., Hassabis, Demis, Platt, John. C.,... Bengio, Yoshua. (2019, 10 junio). *Tackling Climate Change with Machine Learning*. arXiv.org. https://arxiv.org/abs/1906.05433

Sanders, Barrie At the Intersection of Climate Change, AI, and Human Rights Law: Towards a Solidarity-Based Approach (Part 1). (s. f.-b). Afronomicslaw.org, 17 November 2023 https://www.afronomicslaw.org/category/analysis/intersection-climate-change-ai-and-human-rights-law-towards-solidarity-based

School, S. L. (s. f.-a). *EU Artificial Intelligence Act: The European Approach to AI | Stanford Law School*. Stanford Law School. https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/

School, S. L. (s. f.-b). *EU Artificial Intelligence Act: The European Approach to AI | Stanford Law School*. Stanford Law School. https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/

Shaping Europe's digital future. (s. f.). European Commission. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/shaping-europes-digital-future_en

The EU AI Act and environmental protection: the case for a missed opportunity | Heinrich Böll Stiftung | Brussels office - European Union. (2024, 8 abril). Heinrich Böll Stiftung | Brussels Office - European Union. https://eu.boell.org/en/2024/04/08/eu-ai-act-missed-opportunity

Wu, Carole-Jean, Raghavendra, Ramya, Gupta, Udit, Acun, Blige, Maeng, Kiwan, Chang, Gloria, Behram, Fiona Aga, Huang, James, Bai, Charles, Gschwind, Michael, Gupta, Anurang, Ott, Myle, Melnikov, Anastasia, Candido, Salvatore, Brooks, David, Chauhan, Greta, Lee, Benjamin., Lee, Hsien-Hsin S., Akyildiz, Bugra.,... Hazelwood, Kim. (2021, 30 octubre). *Sustainable AI: Environmental Implications, Challenges and Opportunities*. arXiv.org. https://arxiv.org/abs/2111.00364

Capítulo 13

REPENSAR LA PARTICIPACIÓN CIUDADANA ANTE LA JUSTICIA CONSTITUCIONAL: UN ANÁLISIS DESDE EL CONSTITUCIONALISMO DELIBERATIVO INCLUSIVO Y LA INTELIGENCIA ARTIFICIAL[1]

Alejandro Cortés-Arbeláez
Investigador Predoctoral
Universitat Pompeu Fabra

RESUMEN: Este artículo explora una concepción inclusiva del constitucionalismo deliberativo (CD) con un enfoque en la participación ciudadana –mediada por tecnologías digitales basadas en inteligencia artificial (TDIA)– en la justicia constitucional. En la introducción, se contextualiza el CD como una teoría consolidada en el debate sobre la relación entre revisión judicial y democracia. A continuación, se caracteriza al CD a través de conceptos como el "continuo deliberativo" y los "espacios constitucionales deliberativos", y se revisan diversas posturas teóricas que abogan por diferentes grados de participación ciudadana en la interpretación constitucional, destacando las contribuciones de Rawls, Habermas y Lafont. El artículo propone una visión inclusiva del CD que promueve la participación ciudadana en los procesos de revisión judicial. Se argumenta que, aunque las teorías de Rawls y Habermas reconocen la importancia de la ciudadanía en la interpretación constitucional, ambas atribuyen un rol pasivo a los ciudadanos. Se utiliza la teoría participativa de Lafont como punto de partida para elaborar una propuesta más ambiciosa que integre plenamente a los ciudadanos en la justicia constitucional. Así, se explica cómo las TDIA pueden transformar la justicia constitucional mediante tres mecanismos: (i) *información*, a través de tribunales constitucionales ampliados en línea que faciliten el acceso y comprensión de los procesos judiciales;

1 Este artículo se deriva de la tesis doctoral que actualmente estoy escribiendo en el Grupo de Investigación de Filosofía del Derecho de la Universidad Pompeu Fabra (UPF). Agradezco a mis supervisores, José Luis Martí y Sebastián Linares, por su apoyo y guía en este proceso. He presentado versiones previas de este trabajo en el Seminario de Doctorandos del Grupo de Investigación de Filosofía del Derecho de la UPF, en el Seminario de Profesores de la Escuela de Derecho de la Universidad EAFIT (Colombia) y en la Conferencia Anual de la International Society of Public Law (ICON•S) que tuvo lugar en julio de 2024. Agradezco a los participantes de estos eventos por sus comentarios y sugerencias. Este trabajo ha sido apoyado por la subvención CEX2021-001169-M (financiada por el MICIU/AEI/10.13039/501100011033).

(ii) *deliberación*, mediante plataformas de deliberación online que permitan una participación ciudadana masiva y deliberativa en tribunales constitucionales; y (iii) *colaboración*, a través de ejercicios colaborativos de interpretación constitucional que involucren a expertos y ciudadanos. En la última sección se recapitula y presentan algunas conclusiones.

PALABRAS CLAVE: Constitucionalismo deliberativo, revisión judicial, justicia constitucional, democracia, deliberación, participación ciudadana, inteligencia artificial

1. INTRODUCCIÓN

En los últimos diez años, el *constitucionalismo deliberativo* (CD) se ha consolidado como una de las teorías más sólidas para desarrollar la clásica discusión de teoría constitucional sobre la *tensión entre revisión judicial y democracia*. El CD es un *concepto* dentro del cual caben diversas *concepciones*[2], por lo cual cualquier propuesta que pretenda basarse en esta teoría debe explicar con precisión tanto el concepto como las distintas concepciones del CD, con el fin de evitar malentendidos e imprecisiones conceptuales.

En este artículo, siento las bases para el desarrollo de una *concepción inclusiva del CD* específicamente enfocada en la justicia constitucional y soportada en el uso de tecnologías digitales basadas en inteligencia artificial (TDIA) para promover la participación ciudadana ante tribunales constitucionales[3]. En la segunda sección **(2)** de este trabajo desarrollo esta idea en términos generales y en la tercera sección **(3)** recapitulo y presento algunas conclusiones.

2. UNA CONCEPCIÓN INCLUSIVA DEL CONSTITUCIONALISMO DELIBERATIVO: REVISIÓN JUDICIAL PARTICIPATIVA MEDIADA POR TECNOLOGÍAS DIGITALES BASADAS EN INTELIGENCIA ARTIFICIAL

Esta sección está dividida en tres subsecciones. En la subsección 2.1. presento una caracterización *parcial* del CD[4] recurriendo a dos conceptos que denomino el "continuo deli-

2 Respecto de la diferencia entre *conceptos* y *concepciones*, González Ricoy y Queralt señalan lo siguiente: "Según esta distinción, los conceptos sólo incluirían los elementos que antes llamábamos centrales. Serían, por tanto, más sucintos y menos controvertidos que las concepciones, que serían interpretaciones completas y más disputadas de los conceptos. Así, el concepto de discriminación podría definirse como «trato desfavorable por pertenencia a un grupo», mientras que sus diversas concepciones detallarían qué tratos, hacia qué grupos y con qué efectos cuentan como desfavorables. O el concepto de democracia podría definirse como «gobierno del pueblo», mientras que sus diversas concepciones detallarían quiénes forman el pueblo, cómo gobiernan y sobre qué. Los conceptos serían, según Timothy Williamson, sucintos como las entradas de un diccionario, mientras que las concepciones serían extensas como las de una enciclopedia. De modo que las controversias entre concepciones reñidas de la discriminación o la democracia serían compatibles con el acuerdo sobre tales conceptos, cuyo significado de mínimos, al ser compartido, sería más apto para el análisis conceptual" (González Ricoy y Queralt 2021, 31-32).

3 En este texto utilizo el término "tribunal constitucional" de manera abstracta para referirme tanto a tribunales constitucionales *stricto sensu* como a tribunales supremos que cumplen la función de revisión de constitucionalidad de la legislación.

4 La caracterización que presento es parcial debido a que me limito a poner sobre la mesa un marco heurístico que permite clasificar las distintas posturas que, dentro del campo disciplinar del CD, se pueden asumir frente al debate sobre la tensión entre revisión judicial y democracia; una caracterización integral del CD necesitaría ir más allá y analizar, desde una perspectiva sistémica, asuntos relacionados con otros elementos del sistema constitucional, como por ejemplo los mecanismos de reforma y reemplazo constitucional y el rol de mecanismos de democracia directa como los referendos, entre otros. Agradezco a Vicente Aylwin por haberme llamado la atención sobre la necesidad de hacer esta precisión. Para un análisis sucinto, pero completo, del CD, véase: Kong y Levy (2018). Un estudio extenso y detallado del CD se encuentra en Levy et al. (2018)

berativo" y los "espacios constitucionales deliberativos". En la subsección 2.2., basándome en los trabajos de Rawls (1996), Habermas (1998) y Lafont (2020), postulo y defiendo una *concepción inclusiva* del CD en la cual cobra importancia la idea de abrir las puertas de la justicia constitucional a la participación ciudadana y se materializa la propuesta de una *revisión judicial participativa*. En la subsección 2.3. discuto cómo las TDIA pueden utilizarse para acercarse parcialmente al ideal regulativo[5] de la revisión judicial participativa en la concepción inclusiva del CD.

2.1. Una caracterización del campo disciplinar del constitucionalismo deliberativo: el continuo deliberativo y los espacios constitucionales deliberativos

Caracterizo al CD como una teoría política que en los últimos diez años ha pasado de ser "un subcampo de la democracia deliberativa" (Levy 2018, 351) a convertirse progresivamente en un campo disciplinar autónomo cuyo objeto de estudio es la *deliberación en espacios constitucionales* (Giuffré, 2024, 18)[6]. Aunque lo que une a los constitucionalistas deliberativos es su preocupación común por la deliberación en espacios constitucionales, lo que les separa a unos de otros y produce divisiones internas dentro del CD como campo disciplinar es la delimitación de lo que se puede y debe considerar un espacio constitucional en donde pueden desarrollarse diálogos deliberativos; esto es, la delimitación del *universo de los espacios constitucionales deliberativos*.

Es posible organizar las distintas posturas en esta discusión interna dentro del CD mediante una figura heurística que denomino el *continuo deliberativo.* De acuerdo con esta, la deliberación es una variable continua que puede presentarse con distintos grados de intensidad dentro de un sistema institucional; se trata de una cuestión de grado, y no de todo o nada. Pero, aunque se trate de una cuestión gradual, no puede asumir cualquier valor: el continuo deliberativo se enmarca en un intervalo que tiene dos límites a cada lado.

Para explicar cuáles son estos límites, voy a recurrir a dos jueces constitucionales hipotéticos que representan los puntos extremos del continuo deliberativo[7]. En tanto

5 Un ideal regulativo puede entenderse como un "*estado de cosas* que evaluamos como *deseable o correcto* [cursivas en el original]", incluso si juzgamos que este es inalcanzable en la práctica (Martí 2006, 25).

6 Esta teoría política tiene variaciones tanto normativas como empíricas, pues mientras algunos autores están preocupados por asuntos como la justificación de prácticas deliberativas en espacios constitucionales (Bello Hutt 2020), otros están enfocados en asuntos como el análisis de prácticas de deliberación en espacios constitucionales específicos (Arguelhes et al. 2024; Silva 2018). Es importante aclarar que buena parte de la literatura reciente de orientación normativa que se puede inscribir dentro del constitucionalismo deliberativo está *contextual e institucionalmente situada*, pues desarrolla discusiones normativas a partir del análisis y discusión de arreglos e innovaciones institucionales que buscan, precisamente, promover la deliberación en espacios constitucionales (Gargarella 2016, 2019b, 2019a).

7 Algunas personas me han señalado que resulta paradójico que recurra a la figura de jueces constitucionales hipotéticos en un trabajo que aboga justamente por poner a los ciudadanos, y no exclusivamente a los jueces, en

que figuras ficticias, estos personajes son meramente ilustrativos y no se corresponden con algún autor o posición real, pues únicamente buscan mostrar las *fronteras posibles del constitucionalismo deliberativo* en el debate sobre la relación entre revisión judicial y democracia.

A la primera la llamo el *juez elitista epistémico*. Se trata de un juez constitucional que considera que la interpretación constitucional es una actividad que le corresponde única y exclusivamente a él y sus compañeros de tribunal, en virtud de sus superiores capacidades epistémicas en materia de interpretación constitucional. Por ello, para el *juez elitista epistémico* el tribunal constitucional no debe ni siquiera preocuparse por escuchar los argumentos de otras ramas del poder público, y muchísimo menos de la ciudadanía. De manera coherente con su postura, este juez constitucional considera que únicamente los argumentos estrictamente lógico-racionales emitidos *al interior* del tribunal tienen valor para los procedimientos constitucionales-deliberativos.

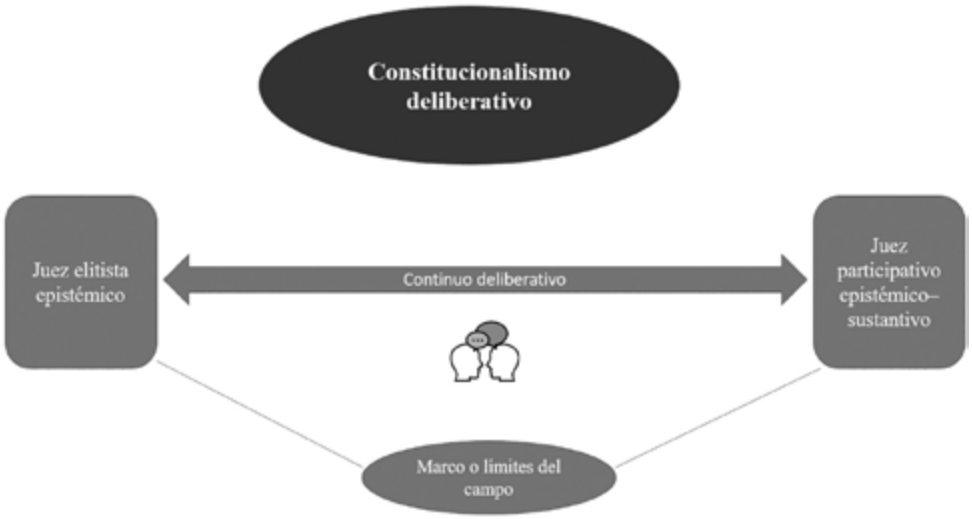
A la segunda la llamo el *juez participativo epistémico–sustantivo*. Este es un juez constitucional que considera que la interpretación constitucional es una tarea que corresponde realizar a *todos los potenciales afectados por las decisiones constitucionales*. Para este personaje, cualquier argumento expresado por cualquier persona debe tener, en principio, igual consideración en las conversaciones constitucionales, siempre que provenga de un potencial afectado por la decisión constitucional objeto de discusión que realice un esfuerzo genuino por tomar parte en el ejercicio de deliberación constitucional. En razón de lo anterior, el *juez participativo epistémico–sustantivo* considera que la justicia constitucional debe abrirse en la mayor medida posible, no solamente al diálogo con otras ramas del poder público, sino especialmente a la participación ciudadana. En línea con su perspectiva, este juez constitucional considera que intervenciones que en sentido estricto no parecen ejercicios de deliberación lógico-racional, como aquellas propias de la protesta social, *pueden* tener valor epistémico para los procedimientos deliberativos porque, por ejemplo, pueden interpretarse como formas de expresión de grupos desaventajados cuyas voces se encuentras excluidas de los procesos de toma de decisiones[8].

el centro de los procesos de justicia constitucional. Entiendo la crítica y no me cierro a cambiar estas figuras en el futuro por, por ejemplo, ciudadanos "de a pie" hipotéticos. Sin embargo, estas figuras son solamente un recurso heurístico al cual me parece útil apelar para ilustrar una discusión relacionada con la justicia constitucional, y no representan un compromiso con una perspectiva juricéntrica "que se enfoca exclusivamente en el funcionamiento interno de las cortes sin prestar suficiente atención al sistema político en el cual estas operan y en donde juegan roles institucionales específicos". Sobre los problemas de esta perspectiva juricéntrica, véase: Lafont (2020, 225).

8 Así, implícita o explícitamente, el *juez participativo epistémico–sustantivo* asume una visión sistémica de la democracia deliberativa. Sobre la aproximación sistémica a la democracia deliberativa, véase: Mansbridge (1999), Mansbridge et al. (2012) y Owen y Smith (2015). Sobre análisis sistémicos aplicados al constitucionalismo deliberativo y la revisión judicial, véase: Bello Hutt (2017) y Valentini (2022). Incluso dejando de lado las razones epistémicas, el *juez participativo epistémico–sustantivo* considera que existe una justificación basada en valores sustantivos, como la igual autonomía, la igualdad política y el respeto mutuo, para la inclusión de la ciudadanía en las conversaciones constitucionales. En este sentido, el *juez participativo epistémico–sustantivo* asume como suya una justificación mixta, tanto epistémica como sustantiva, de la democracia deliberativa (Martí 2006, 213).

Así, dentro del continuo deliberativo delimitado por el *juez elitista epistémico* y el *juez participativo epistémico-sustantivo*[9], algunos abogan por restringir, y otros por ampliar, los actores sociales e institucionales que deben participar de las discusiones constitucionales. Pero en todo caso, el requisito mínimo para ser parte del campo del constitucionalismo deliberativo es el compromiso subyacente con el ideal regulativo de la democracia deliberativa[10]. En el siguiente esquema resumo estas ideas gráficamente.

Esquema 1. El constitucionalismo deliberativo como campo disciplinar



Fuente: Elaboración propia

9 Estas figuras ficticias, como dije, representan las fronteras posibles del constitucionalismo deliberativo. Entre este par de extremos ficticios, se pueden encontrar muchas posiciones y variaciones reales. Pero los integrantes del campo disciplinar estamos unidos por una especie de consenso superpuesto cuyo marco es el *continuo deliberativo*: el acuerdo sobre que las decisiones constitucionales se tienen que basar en procedimientos deliberativos. De esta manera, en la discusión sobre la tensión entre revisión judicial y democracia, el CD, en lugar de "centrarse en la cuestión de si la revisión judicial es ilegítima porque frustra la voluntad democrática", afirma que "la revisión judicial es legítima en la medida en que facilita la deliberación democrática, tanto dentro de las instituciones del poder público (incluidos los tribunales) como en la sociedad en general" (Kong y Levy 2018, 634).

10 Sobre democracia deliberativa, los siguientes textos son imprescindibles: Bächtiger et al. (2018) y Martí (2006). Es importante aclarar que existen distintas concepciones de la democracia deliberativa. Personalmente, me identifico con una concepción participativa de la democracia deliberativa muy cercana a la esbozada por Lafont, aunque tal vez con cierta inclinación neo-republicana, que considera que la democracia deliberativa tiene justificaciones normativas tanto epistémicas como sustantivas. Al respecto, véase: Cortés-Arbeláez (2023).

Como señalé en el párrafo anterior, dentro del campo disciplinar del CD tal como lo reconstruí aquí caben diversas perspectivas. En un trabajo reciente, Giuffré (2023d) propuso una taxonomía de cuatro variantes del CD, cuyo criterio de clasificación es, precisamente, el número y tipo de actores e instituciones que deben participar en las deliberaciones constitucionales. O, en los términos en que lo he venido planteando en este trabajo, el número y tipo de *espacios constitucionales deliberativos* en los cuales tienen lugar las conversaciones que, en últimas, dan forma al ejercicio de interpretación constitucional.

En primer lugar, se encuentran las teorías del diálogo intrajudicial, que consideran que el espacio constitucional deliberativo se restringe al tribunal constitucional. En segundo lugar, están las teorías del diálogo transjudicial, para las que los espacios constitucionales deliberativos son aquellos donde interactúan tribunales nacionales e internacionales, que afinan sus argumentos y perspectivas mediante la deliberación entre jueces de distintas cortes. En tercer lugar, es posible ubicar las teorías del diálogo interinstitucional, que consideran que los espacios constitucionales deliberativos no son exclusivamente judiciales, por lo cual abogan por la deliberación de los jueces constitucionales con otras ramas del poder público, especialmente la legislativa. En cuarto lugar, y de especial importancia para este artículo, tenemos a las *teorías del diálogo inclusivo*, que sostienen que los espacios constitucionales deliberativos incluyen a los tribunales, otras ramas del poder público y *a la sociedad*. En consecuencia, nos dicen las teorías del dialogo inclusivo, *los ciudadanos deben jugar un rol fundamental en el ejercicio de interpretación constitucional* (Giuffré 2023d)[11].

2.2. Una concepción inclusiva del constitucionalismo deliberativo y la idea de la revisión judicial participativa

Rawls y Habermas, probablemente los dos filósofos políticos más influyentes de los últimos 50 años, sostienen posiciones que resultan de utilidad para defender una *concepción inclusiva* del CD que apuesta por la necesidad de involucrar activamente a los ciudadanos en la interpretación constitucional mediante el impulso de la participación ciudadana ante la justicia constitucional. Estos autores, como veremos, esgrimen argumentos que apoyan la idea de que la ciudadanía es y debe ser un actor central en la interpretación constitucional. Sin embargo, ambos atribuyen un rol pasivo a los ciudadanos en la interpretación constitucional mediante procesos de revisión judicial ante tribunales constitucionales. Por ello, si bien resalto el potencial de sus argumentos para defender una concepción inclusiva del CD, me distancio de ellos precisamente debido a que, en materia de diseño institucional, terminan asumiendo una posición menos inclusiva de la que sugieren sus propios compromisos normativos[12].

11 En los términos del continuo deliberativo y las figuras ficticias de las que hablé, al extremo izquierdo, cerca del *juez elitista epistémico*, se encuentra la versión más excluyente (no lo digo en sentido despectivo, sino analítico) de las teorías del dialogo intrajudicial, mientras que, al extremo derecho, cerca *del juez participativo epistémico*, se encuentra la versión más incluyente (no lo digo en sentido de elogio, sino analítico) de las teorías del diálogo inclusivo.

12 Es importante señalar que las ideas de Rawls y Habermas sobre la participación ciudadana son mucho más amplias y complejas de lo que aquí se menciona. Esta sección intenta resaltar algunos aspectos específicos de

En *El liberalismo político,* Rawls (1996) reconoce que la interpretación constitucional es una tarea que no corresponde de manera exclusiva al tribunal constitucional, sino a todas las ramas del poder público y a la sociedad en conjunto. "La constitución no es lo que el tribunal supremo dice que es. Es, antes bien, lo que el pueblo, actuando constitucionalmente a través de las otras ramas, permite eventualmente al tribunal supremo decir que es. Una particular interpretación de la constitución puede ser impuesta al tribunal mediante enmiendas, o a través de un amplia y continuada mayoría política, como ocurrió en el New Deal en Estados Unidos" (Rawls 1996, 273).

A pesar del tono radicalmente democrático de su afirmación, Rawls termina suscribiendo una concepción no-inclusiva del CD en la cual la ciudadanía pasa a jugar un rol secundario en los procesos de revisión judicial ante la justicia constitucional (Giuffré 2023a, 2023d)[13]. De esta manera, aunque a la interpretación rawlsiana subyace el potencial para desarrollar una concepción inclusiva del CD, Rawls termina decantándose por una teoría del diálogo intrajudicial en la cual "el tribunal supremo está llamado a ser el centro de la controversia" (Rawls 1996, 274-75), mientras que los ciudadanos son relegados a la periferia del sistema político (Giuffré 2023d, párr. 8).

Habermas (1998), por su parte, desarrolla en *Facticidad y validez* una interpretación procedimental de la revisión judicial inspirada en buena parte por la propuesta planteada por Ely (1980) según la cual la tarea de la justicia constitucional consiste en proteger, no los valores sustantivos plasmados en el texto constitucional, sino las condiciones procedimentales que hacen posible el funcionamiento del proceso democrático[14]. Sin embargo, Habermas se distancia de Ely en lo que se refiere a la concepción subyacente de democracia que se asume como trasfondo del argumento: mientras Ely entiende a la democracia en términos pluralis-

sus argumentos, sin pretender abarcar toda la riqueza y profundidad de sus aportes teóricos. Por cuestiones de espacio, me concentro en ciertos puntos clave relevantes para el tema, aunque soy consciente de que sus obras ofrecen muchos otros matices que aquí no puedo explorar. Agradezco a Roberto Gargarella por recordarme la importancia de matizar este análisis y de evitar, en la medida de lo posible, simplificar en exceso el trabajo de pensadores cuyos aportes a la teoría política no pueden resumirse

13 Lo anterior debido a que, para Rawls, el tribunal constitucional es "la única rama del estado que es palmaria y visiblemente" un "modelo de la razón pública", por lo cual su papel institucional consiste en "hacer que sus razones sean consistentes y adecuarlas a una visión constitucional coherente que abarque el espectro global de sus decisiones", lo que significa que "es tarea de los jueces intentar desarrollar y expresar, en sus opiniones razonadas, la mejor interpretación de la constitución que puedan usando su conocimiento de lo que exigen la constitución y los precedentes constitucionales" (Rawls 1996, 271).

14 "[L]a Constitución, en nuestras condiciones de pluralismo social y cultural, tampoco puede entenderse como un orden jurídico global de tipo concreto que impusiese *a priori* a la sociedad una determinada forma de vida. Antes la Constitución fija los procedimientos políticos conforme a los que los ciudadanos, ejercitando su derecho de autodeterminación, pueden perseguir cooperativamente y con perspectivas de éxito el proyecto de establecer formas justas de vida […] Sólo las *condiciones procedimentales de la génesis democrática de las leyes* aseguran la legitimidad del derecho establecido. Partiendo de esta comprensión democrática de fondo, también cabe dar a las competencias del Tribunal Constitucional un sentido que responda a la intención a que en el Estado de derecho responde la división de poderes: el Tribunal Constitucional habría de proteger precisamente ese sistema de los derechos que posibilita la autonomía privada y pública de los ciudadanos" (Habermas 1998, 336).

tas, Habermas lo hace en términos de una democracia deliberativa de "doble vía" en la cual resulta fundamental que las decisiones respaldadas por el poder coercitivo del Estado –el "poder administrativo"– respondan a los procesos informales de formación de la opinión en la sociedad civil –el "poder comunicativo" (Velasco 2003, 106-14). "Desde esta perspectiva, el Tribunal Constitucional ha de operar dentro del marco de sus competencias en el sentido de que el proceso de producción de normas se efectúe en las condiciones de una *política deliberativa*, que son las que fundan legitimidad" (Habermas 1998, 348).

La interpretación deliberativo-procedimental que Habermas hace de la revisión judicial se enmarca en una concepción deliberativa de la democracia especialmente preocupada por la necesidad de establecer canales de conexión y comunicación entre "la formación política de la opinión en circuitos informales de comunicación política" y "la formación de la voluntad en los organismos parlamentarios" (Habermas 1998, 349), y por tanto particularmente comprometida con la defensa de una sociedad civil activa y participativa. Sin embargo, en lo que se refiere al diseño institucional de la justicia constitucional, Habermas no parece especialmente interesado en defender la idea de que sea necesario establecer canales de conexión y comunicación entre "la formación política de la opinión en circuitos informales de comunicación política" y el tribunal constitucional.

Lo anterior debido a que, desde su punto de vista, el tribunal constitucional es una institución en la cual tienen lugar los *discursos jurídicos de aplicación* del derecho, pero no los *discursos políticos de fundamentación del derecho*[15], más propios estos últimos del ámbito parlamentario[16] (Zurn 2007, 246). Es cierto que en el capítulo final de *Facticidad y validez*

15 La diferencia entre discursos políticos de fundamentación propios del ámbito parlamentario y los discursos jurídicos de fundamentación propios del ámbito judicial responde a las diferentes lógicas de la argumentación correspondientes a cada ámbito. "La competencia legislativa, que por principio corresponde a los ciudadanos en su totalidad, se encargan de desempeñarla cuerpos parlamentarios, que *fundamentan* leyes conforme a un procedimiento democrático. Las leyes constituyen la base de las pretensiones jurídicas individuales; éstas resultan de la aplicación de las leyes a casos particulares, ya vengan implementadas esas leyes autoejecutivamente, ya lo vengan por vía administrativa […] [L]a diferencia que, en lo tocante a lógica de la argumentación, se da entre fundamentación de normas y aplicación de normas, se refleja en las formas de comunicación que representan respectivamente los discursos de fundamentación y de aplicación, los cuales, por tanto, habían de institucionalizarse de modo distinto. En los discursos jurídicos de aplicación ha de tomarse una decisión acerca de cuál de las normas presupuestas como válidas es la que se ajusta a una situación descrita de la forma más completa posible en todos sus rasgos relevantes. Este tipo de discurso exige una constelación de roles en la que las partes (o, en su caso, las autoridades estatales encargadas de hacer las correspondientes diligencias) puedan presentar todos los aspectos controvertidos de un caso ante un juez como representante de una comunidad jurídica encargada de juzgar imparcialmente, y además una distribución de competencias conforme a la que el tribunal ha de justificar su sentencia ante un espacio público jurídico ilimitado en principio. En cambio, en los discursos de fundamentación sólo hay en principio participantes. Por otro lado, la administración de justicia, para imponer sus decisiones –y ejecutar las sentencias– ha de recurrir a los medios de represión del aparato estatal y dispone, por tanto, ella misma de poder administrativo. Por esta razón la justicia ha de quedar separada del poder legislativo, impidiéndosele así que se programe a sí misma. Así se explica el *principio de la vinculación de la justicia al derecho vigente*" (Habermas 1998, 240-41).

16 Es importante aclarar que, en todo caso, Habermas sí considera opciones de institucionalización de la función de revisión judicial diferentes a la asignación de la misma a un tribunal constitucional. De manera más específica, el autor señala que la posibilidad de asignar la función de revisión abstracta de normas a una comisión

Habermas reconoce que en la justicia constitucional "los discursos jurídicos de aplicación han de complementarse de forma clara y reconocible con elementos tomados de los discursos de fundamentación" para concretar el contenido de la legislación, y sugiere que la legitimación de esta "formación cuasi legisladora de la opinión y de la voluntad" requiere de "la institucionalización de un espacio público-jurídico que vaya más allá de la actual cultura de expertos" (Habermas 1998, 525-26). Sin embargo, el autor no va más allá de esta consideración general, por lo cual el potencial de la teoría de la democracia deliberativa habermasiana para promover una concepción inclusiva del CD que se comprometa con una revisión judicial participativa no queda suficientemente desarrollado en su obra.

En *Democracia sin atajos*[17], Lafont (2020) propone una *(re)interpretación participativa de la revisión judicial.* De acuerdo con esta, la revisión judicial no debe considerarse como un "atajo expertocrático", sino más bien como un "iniciador de conversaciones", es decir, una forma política de participación ciudadana que permite a los ciudadanos exigir, tanto a las autoridades como a otros ciudadanos, la justificación de las leyes y políticas, *desencadenando un proceso de impugnación legal a través de la revisión judicial* (Lafont 2020, 238-39). A juicio de Lafont, esta interpretación participativa de la revisión judicial permite que se entienda normativamente a la misma como una "institución de control democrático".

integrada al parlamento merece ser considerada seriamente. "Los tribunales constitucionales cumplen normativamente varias funciones a la vez. Aunque las distintas competencias convergen en la tarea de decidir en última instancia cuestiones de interpretación de la Constitución y, por tanto, en mantener también la coherencia del orden jurídico, el agavillamiento de esas competencias en una única institución, no resulta obligatorio sin más desde el punto de vista de una teoría de la Constitución […] La competencia concerniente a recursos de amparo y al control concreto de normas (es decir, la competencia relativa a los casos en que los tribunales de instancia suspenden temporalmente un proceso para obtener, con motivo de algún asunto concreto, una decisión acerca de la constitucionalidad de la norma que se tiene que aplicar) es la menos problemática desde el punto de vista de la división de poderes. Aquí el Tribunal Constitucional actúa en el sentido de una unificación y coherentización del derecho […] La colisión entre las funciones del Tribunal Constitucional y lo que se dirían habrían de ser funciones del legislador legitimado democráticamente sólo se agudiza en el ámbito del control abstracto de normas […] Y, en todo caso, es digna de tenerse en cuenta la consideración de si la revisión de tal decisión parlamentaria no podría efectuarse a su vez en forma de un autocontrol del legislador mismo, organizando tal autocontrol a modo de tribunal, e institucionalizándolo, por ejemplo, en una comisión parlamentaria compuesta también por profesionales del derecho. Tal internalización de la autorreflexión acerca de las propias decisiones tendría la ventaja de mover al legislador a tener presente desde el principio durante sus propias deliberaciones el contenido normativo de los principios constitucionales. Pues éste tiende a perderse de vista cuando en el tráfago de los negocios parlamentarios las cuestiones éticas y morales quedan *redefinidas* como cuestiones susceptibles de negociarse, es decir, como cuestiones susceptibles de compromiso. En este aspecto, la diferenciación institucional de un procedimiento autorreferencial de control de normas, que fuese competencia del Parlamento, contribuiría quizá a aumentar la racionalidad del proceso de producción legislativa. Y ello resulta tanto más recomendable si, conforme al sentido de nuestro análisis, se parte de que la división de poderes tiene ante todo la finalidad de que la Administración no se autonomice frente al poder generado comunicativamente" (Habermas 1998, 313-14).

17 Cabe señalar que en este libro Lafont desarrolla una concepción de la democracia que puede entenderse como "síntesis de las ideas de Habermas y Rawls, pero llevadas más allá, puestas al servicio de una democracia fuerte y participativa, un ideal exigente para el que no hay atajos posibles" (Martí 2023, 88).

A pesar de que la interpretación participativa de la revisión judicial elaborada por Lafont es normativamente atractiva, resulta ser insuficientemente ambiciosa desde una perspectiva que se tome en serio *el valor de la participación ciudadana en los procesos de revisión judicial*. En consecuencia, se trata de una propuesta que no satisface los exigentes estándares normativos de la concepción inclusiva del CD. Aunque existen varias razones para sostener lo anterior (Chambers 2020; Gargarella 2023; Giuffré 2022, 2023b; Linares 2023), aquí me quiero centrar en una: en el esquema de Lafont, el rol de los ciudadanos en la revisión judicial es meramente "contestatario". Pueden, sí, *iniciar* una conversación constitucional. Pero una vez iniciada, los ciudadanos parecen quedar limitados al rol de espectadores pasivos, pues la conversación es absorbida por el tribunal constitucional y el parlamento (Gargarella 2023, 137).

Esto no significa que la teoría de la revisión judicial de Lafont deba ser desechada. Por el contrario, esta es una contribución de peso al campo del CD que tiene un potencial significativo para revitalizar las discusiones sobre cómo hacer más democrática a la justicia constitucional. La tarea consiste, más bien, en utilizar la teoría de Lafont como punto de partida para la elaboración de una teoría de la revisión judicial no solamente más comprometida con poner a los ciudadanos genuinamente en el centro de los procesos de justicia constitucional, sino también más sensible respecto de las condiciones institucionales que se requieren para ello (Cortés-Arbeláez, 2023, 70-79). En la siguiente subsección siento las bases para el desarrollo de una teoría de la revisión judicial participativa mediada por TDIA.

2.3. Una revisión judicial participativa impulsada por tecnologías digitales basadas en inteligencia artificial

En una contribución reciente, Martí analizó los posibles roles de las tecnologías digitales en la democracia deliberativa (Martí 2021). Planteó tres formas en las que las tecnologías digitales pueden ayudar a aumentar nuestra inteligencia individual y colectiva y, de esta manera, fortalecer la democracia deliberativa:

I. Información. En primer lugar, pueden ayudarnos a resolver el problema de la "economía de la información" haciendo "posible por primera vez la divulgación plena y la transparencia total a un coste muy bajo" (Martí 2021, 215), permitiendo así a los ciudadanos estar informados de las acciones gubernamentales, y a los gobiernos acceder a las aportaciones de la ciudadanía casi en tiempo real.

II. Deliberación. En segundo lugar, pueden mejorar el poder de la deliberación con herramientas de deliberación online que hagan uso de mecanismos de inteligencia artificial, como el aprendizaje automático, para estructurar mejor las interacciones argumentativas ayudándonos a "identificar quién está presentando argumentos similares en discusiones separadas y qué argumentos reúnen más apoyo y emergen como más fuertes que otros" (Martí 2021, 218). En un sentido similar, Noveck ha argumentado que si "las instituciones utilizaran el aprendizaje automático para resu-

mir y analizar los comentarios, podrían comprender mejor la participación pública y aumentar el valor epistémico del compromiso" (Noveck 2021, 129).

III. Colaboración. En tercer lugar, pueden incentivar la colaboración humana mediante mecanismos de inteligencia colectiva potenciados por la tecnología (Martí 2021, 218). Uno de los ejemplos más relevantes de esto es la idea de CrowdLaw, un concepto que ha sido desarrollado, entre otros, por Noveck y Martí. CrowdLaw se refiere a una forma de participación ciudadana directa basada en la tecnología que "está conceptualmente conectada con la elaboración de leyes y la toma de decisiones públicas de cualquier tipo y a cualquier nivel, desde la elaboración de constituciones hasta la legislación, la formulación de políticas y la toma de decisiones judiciales" (Alsina y Martí 2018, 318). CrowdLaw se diferencia de formas anteriores de mecanismos de participación pública, entre otras cosas, porque está institucionalizado y busca obtener experiencia e ideas de la ciudadanía, y no simplemente opiniones (Noveck 2018, 366). Se mueve, pues, no sólo por objetivos estrictamente participativos, sino también por ambiciones epistémicas.

Siguiendo este esquema, a continuación, planteo una idea sobre cómo insertar TDIA en procesos de revisión judicial de manera tal que, en línea con la concepción inclusiva del CD, la justicia constitucional se tome en serio la participación ciudadana y sea posible materializar una *revisión judicial participativa*.

2.3.1. Información: un tribunal constitucional ampliado y en línea

En *Online Courts and the Future of Justice*, Susskind (2019) propone la creación de *tribunales en línea*. Estos están pensados para la justicia civil, pero el autor considera que pueden extenderse a otras áreas del derecho. En su propuesta, estos tribunales en línea deben cumplir dos funciones: (a) juicios en línea (*online judging*) y (b) servicios judiciales ampliados/tribunales ampliados (*extended court services/extended courts*). La primera categoría hace referencia principalmente al desarrollo de procesos judiciales asincrónicos en los cuales "los jueces escuchan los argumentos y las pruebas, adoptan sus decisiones y las dan a conocer a las partes y al mundo en general sin pisar una sala física" (Susskind 2019, 143). La segunda categoría alude a la incorporación en la estructura institucional de los tribunales en línea de herramientas para la prestación de servicios que "incluyen herramientas para ayudar a los usuarios a comprender sus derechos, deberes y opciones que se les ofrecen, instalaciones que ayudan a los litigantes a reunir sus pruebas y formular sus argumentos, y sistemas que asesoran sobre acuerdos extrajudiciales o los propician" (Susskind 2019, 61).

Con base en este planteamiento, es plausible plantear el desarrollo de *extended constitutional court services*, que faciliten la interacción de tribunales constitucionales con la ciudadanía. No sería un disparate imaginarse un tribunal constitucional que, como parte de sus funciones y

servicios, cuenta con una plataforma virtual interactiva y amable con el usuario[18] mediante la cual ciudadanos comunes y corrientes pueden informarse sobre asuntos tan básicos como en qué consiste la revisión judicial de constitucionalidad, hasta sobre cómo iniciar procesos de revisión judicial de constitucionalidad ante el tribunal[19].

De esta manera, para la ciudadanía se podrían ver considerablemente reducidos los costos de información necesarios para realizar una acción técnicamente compleja, como el inicio de un proceso de revisión judicial de constitucionalidad. Simultáneamente, el tribunal constitucional contaría con canales que le permitirían conocer las aportaciones y preocupaciones de ciudadanos comunes y corrientes en tiempo real. Aquí, por ejemplo, podría pensarse en que los magistrados del tribunal reciban reportes mensuales automatizados en los cuales se informe sobre cuáles son los asuntos de mayor interés para la ciudadanía, cuáles son las leyes que más inquietud están generando, entre otros.

2.3.2. Deliberación: procesos de justicia constitucional basados en la deliberación ciudadana online

Una vez iniciado un proceso de revisión judicial de constitucionalidad, el tribunal constitucional podría recurrir a tecnologías digitales para fomentar *espacios constitucionales deliberativos* a gran escala que permitan que la ciudadanía tome parte de las deliberaciones que tienen lugar en el tribunal constitucional. Tribunales constitucionales como el colombiano, argentino, brasilero y mexicano ya han hecho esfuerzos importantes por abrir sus deliberaciones hacia la ciudadanía mediante mecanismos como las audiencias

18 Creada y actualizada con herramientas de *diseño legal.* Sobre este concepto, véase: Chung y Kim (2023).

19 Un *extended constitutional court service* podría tener diferentes etapas o niveles. Así, por ejemplo, en un primer nivel podría limitarse a proveer información sobre asuntos básicos como explicar cuál es la naturaleza y funcionamiento de la revisión judicial de constitucionalidad. En este nivel, el ciudadano interactuaría exclusivamente con sistemas de inteligencia artificial, como chatbots. En un segundo nivel, podría tener una herramienta que asesore a ciudadanos que estén pensando en iniciar procesos de revisión judicial de la legislación. Esta podría orientar a los ciudadanos de manera tal que comprendan que la revisión judicial es un arreglo institucional complejo que solamente debe utilizarse cuando existan razones de peso para creer que una ley efectivamente vulnera la constitución, y no para intentar cuestionar legislación que simplemente consideran inadecuada. Adicionalmente, mediante análisis de jurisprudencia basado en herramientas de inteligencia artificial, el *extended constitutional court service* podría interactuar con usuarios que estén interesados en conocer las líneas jurisprudenciales en materia del objeto de interés con el fin de evaluar si los argumentos que ellos consideran que podrían ameritar el inicio de un proceso de revisión judicial efectivamente podrían encontrar receptividad en el tribunal constitucional. En este nivel, el ciudadano interactuaría con sistemas de inteligencia artificial, pero también sería posible solicitar la interacción con funcionarios humanos. En un tercer nivel, el *extended constitutional court service* podría proveer asesoría a los ciudadanos que efectivamente decidan iniciar procesos de revisión judicial, ayudándoles a redactar la acción de inconstitucionalidad en un lenguaje y unos términos que sean eventualmente aceptables para el tribunal constitucional. Adicionalmente, en este mismo nivel el sistema podría, previo consentimiento de los usuarios, sugerir agregar sus acciones de inconstitucionalidad con otras que vayan a ser presentadas por otros usuarios en contra de la misma legislación. En este nivel, el ciudadano interactuaría con sistemas de inteligencia artificial, pero también sería posible solicitar la interacción con funcionarios humanos.

públicas. Estas, de hecho, suelen ser mencionadas en la literatura de constitucionalismo deliberativo como una herramienta importante, aunque usualmente insuficiente, de participación ciudadana ante la justicia constitucional (Gargarella 2019b, 2021, 257-58; Giuffré 2023c, 311).

Este tipo de mecanismos podría fortalecerse mediante TDIA que utilicen plataformas de deliberación online para promover debates de amplia escala ante el tribunal constitucional, en los cuales pueda participar un número genuinamente significativo de ciudadanos. En la literatura sobre deliberación online existen algunas pistas, pensadas para escenarios diferentes a la justicia constitucional, que pueden ser útiles para lograr lo anterior.

Landemore (2023) ha señalado que hay nueve formas en las que la inteligencia artificial puede ayudar a escalar y mejorar procesos de deliberación online: (i) facilitación, (ii) traducción, (iii) comprobación de hechos, y (iv) agrupación de datos. Estas funciones pueden ser realizadas por la inteligencia artificial o por humanos. Pero otras tareas solo pueden ser realizadas por la inteligencia artificial: (v) rastrear todos los intercambios que alguien ha tenido con cualquier otro y medir el grado de solapamiento en el contenido, (vi) medir la calidad de la deliberación, (vii) tomar imágenes cognitivas del grupo, (viii) compartir el consenso entre grupos, y (ix) difundir ideas con alto potencial entre grupos[20].

Con base en innovaciones democráticas digitales soportadas en herramientas de inteligencia artificial, y con el apoyo de plataformas de deliberación online del estilo de *Polis*[21], el tribunal constitucional puede diseñar y adelantar procesos deliberativos que incluyan elementos presenciales y online, con momentos de deliberación tanto sincrónica como asincrónica, y que permitan una genuina deliberación ciudadana a gran escala ante el tribunal constitucional[22].

20 En un sentido similar, Mikhaylovskaya ha argumentado que las innovaciones democráticas digitales pueden contribuir a mejorar tres rasgos de "la deliberación a través de medios digitales, a saber: escala, transparencia y, por último, igualdad en la deliberación" (Mikhaylovskaya 2024, 8). En primer lugar, pueden aumentar la *escala* de las deliberaciones al permitir ejercicios de deliberación online masiva sincrónica y asincrónica. En segundo lugar, sirven para promover la *transparencia* en deliberaciones, pues permiten rastrear cómo se forman los argumentos, las razones proveídas para aceptarlos o rechazarlos, y permiten el uso de herramientas de fácil uso para entender las deliberaciones, como mapas de argumentos y mecanismos de *deliberación visual* (Noveck 2009, 71). En tercer lugar, pueden contribuir a fomentar la *igualdad* en la deliberación al crear asistentes virtuales para los moderadores humanos de deliberaciones online, que les ayuden a rastrear a personas que estén dominando indebidamente las discusiones, grupos de minorías "insulares y discretas" que tengan problemas para hacerse escuchar, y en general distorsiones que introduzcan desigualdades no aceptables en los procesos deliberativos (Mikhaylovskaya 2024, 16).

21 "Polis pretende ofrecer a los ciudadanos una visión dinámica de todo el espectro de opiniones en torno a un tema de debate y se considera una herramienta de democracia directa y deliberativa muy eficaz. Permite al gobierno plantear preguntas políticas al público y, a continuación, utiliza un resumen estadístico para proporcionar información gráfica sobre lo que cree o desea la población en su conjunto. El sistema afirma ser eficaz para lograr el consenso popular en torno a cuestiones polémicas durante un periodo de dos o tres semanas con entre 100 y decenas de miles de participantes o más" (Tsai et al. 2024, 10).

22 Por supuesto, es necesario evitar caer en el tecno-optimismo inmoderado al pensar en propuestas de esta índole. La literatura reconoce que, a pesar de los avances en el desarrollo de herramientas de deliberación online

2.3.3. Colaboración: interpretación constitucional colaborativa

Durante el proceso de revisión judicial, el tribunal constitucional puede complementar la deliberación ciudadana online con ejercicios colaborativos de interpretación constitucional. La idea principal es que, además de crear espacios formales de deliberación en línea para involucrar a una gran cantidad de ciudadanos comunes, lo cual redunda en procesos constitucionales *inclusivos,* el tribunal constitucional puede beneficiarse al abrir espacios para la democracia colaborativa, lo cual redunda en procesos constitucionales *epistémicamente fortalecidos*.

En el paradigma de la democracia colaborativa, que ha sido elaborado principalmente por Noveck (2009), las instituciones ofrecen al público la oportunidad de autoseleccionarse para participar activamente de diferentes maneras en procesos de diseño de política pública. La democracia colaborativa es un enfoque innovador que utiliza la tecnología para mejorar los resultados al solicitar la experiencia (definida ampliamente para incluir tanto el conocimiento científico como la experiencia popular) de ciudadanos autoseleccionados que colaboran en grupos dentro de redes abiertas. Con su experiencia y experticia, los *expertos voluntarios* pueden ampliar los conocimientos de los funcionarios y coordinar sus propias estrategias. Aprovechando el ahorro de costos que ofrece la tecnología, las jerarquías pueden transformarse en ecosistemas de conocimiento colaborativo, cambiando radicalmente la cultura de gobernanza pública (Noveck 2009, 16, 35, 40, 70).

Con base en este paradigma, es posible imaginarse actividades colaborativas en procesos de revisión judicial, que fortalezcan el potencial epistémico de la justicia constitucional mediante la *interpretación constitucional colaborativa*[23] que se desarrollaría, por ejemplo, a través de *amicus curiae* colaborativos en los cuales expertos de distinta índole podrían contribuir de diversas maneras, dependiendo de la temática objeto de discusión.

De esta manera, mediante innovaciones democráticas digitales basadas en inteligencia artificial se podría lograr el aumento de la inteligencia colectiva[24] en los procesos de revisión judicial, de manera tal que se potencie no solamente el carácter inclusivo de la justicia constitucional, sino además su potencial epistémico.

---

potenciadas mediante inteligencia artificial, es necesario no sobreestimar el potencial de estas, pues pueden reproducir sesgos humanos y reforzar la discriminación de personas y grupos ya discriminados (Shortall et al. 2022, 13). Adicionalmente, se corre el riesgo de que con el uso de TDIA para potenciar procesos deliberativos, "la velocidad con la que el grupo encuentra áreas de acuerdo o posibles compromisos podría reducir el debate y la reflexión creativa, genuina y atractiva, ya que el foco se desplaza hacia la eficiencia y el acuerdo en lugar de la exploración de diversas ideas o la práctica de la argumentación constructiva" (Tsai et al. 2024, 12).

23 Para una propuesta de interpretación constitucional colaborativa fundamentada en la teoría del *constitucionalismo popular*, véase: Abat i Ninet (2021, 127).

24 Sobre el uso de TDIA para "aumentar" la inteligencia colectiva de la sociedad, véase: Berditchevskaia y Baeck (2020), Grobbink y Peach (2020) y Pasquale (2020, 13).

3. CONCLUSIONES

En este artículo senté las bases para desarrollar una *concepción inclusiva del CD,* proponiendo la integración de TDIA en tribunales constitucionales para fomentar la participación ciudadana en los procesos de revisión judicial. De esta manera, es posible no solamente hacer más democrática a la justicia constitucional, sino también mejorar su calidad epistémica mediante mecanismos de información, deliberación y colaboración. Mediante la creación de tribunales constitucionales en línea, procesos de deliberación online masiva ante la justicia constitucional y ejercicios de interpretación constitucional colaborativa, es posible imaginar una justicia constitucional abierta a la participación ciudadana. En futuras contribuciones, desarrollaré los detalles específicos de esta propuesta, de la cual este artículo es su esbozo inicial.

4. BIBLIOGRAFÍA

ABAT I NINET, Antoni. *Constitutional Crowdsourcing: Democratising Original and Derived Constituent Power in the Network Society*. Cheltenham, UK Northampton, MA: Edward Elgar Publishing, 2021.

ALSINA, Victòria; MARTÍ, José Luis. "The Birth of the Crowdlaw Movement: Tech-Based Citizen Participation, Legitimacy and the Quality of Lawmaking". *Analyse & Kritik* 40(2) (2018): 317-38.

ARGUELHES, Diego Werneck; CESARIO Alvim, Juliana; NOGUEIRA, Rafaela; WANG, Henrique. "They Don't Let Us Speak: Gender, Collegiality, and Interruptions in Deliberations in the Brazilian Supreme Court". *Journal of Empirical Legal Studies* 21(1) (2024): 174-207. doi:10.1111/jels.12379.

BÄCHTIGER, Andre; DRYZEK, John S.; MANSBRIDGE, Jane; WARREN, Mark. "Deliberative Democracy: An Introduction". En: Bächtiger, Andre; Dryzek, John S.; Mansbridge, Jane; Warren, Mark (eds.): *The Oxford Handbook of Deliberative Democracy*. Oxford University Press, 2018, pp. 1-32.

BELLO HUTT, Donald. "Deliberation and Courts: The Role of the Judiciary in a Deliberative System". *Theoria* 64(3) (2017): 77-103. doi:10.3167/th.2017.6415204.

BELLO HUTT, Donald. "The deliberative constitutionalism debate and a republican way forward". *Jurisprudence. An International Journal of Legal and Political Thought* 12(1) (2020): 69-88.

BERDITCHEVSKAIA, Aleks; BAECK, Peter. *The Future of Minds and Machines: How artificial intelligence can enhance collective intelligence*. Nesta, 2020. https://www.nesta.org.uk/report/future-minds-and-machines/.

CHAMBERS, Simone. "Citizens Without Robes: On the Deliberative Potential of Everyday Politics". *Journal of Deliberative Democracy* 16(2) (2020): 73-80. doi:https://doi.org/10.16997/jdd.388.

CHUNG, Sunghoon; KIM, Jieun. "Systematic Literature Review of Legal Design: Concepts, Processes, and Methods". *The Design Journal* 26(3) (2023): 399-416. doi:10.1080/14606925.2022.2144549.

CORTÉS-ARBELÁEZ, Alejandro. "Judges without Robes: A Republican Approach to Participatory Judicial Review". *Revista Derecho del Estado* (57) (2023): 41-83. doi:10.18601/01229893.n57.03.

ELY, John Hart. *Democracy and Distrust: A Theory of Judicial Review*. Harvard University Press, 1980.

GARGARELLA, Roberto. "Scope and Limits of Dialogic Constitutionalism". En: Bustamante, Thomas; Gonçalves Fernandes, Bernardo (eds.): *Democratizing Constitutional Law. Perspectives on Legal Theory and the Legitimacy of Constitutionalism*, Law and Philosophy Library, Springer, 2016, pp. 119-46.

GARGARELLA, Roberto. "La revisión judicial para las democracias latinoamericanas". En: Niembro, Roberto; Verdugo, Sergio (eds.): *La justicia constitucional en tiempos de cambio*. Suprema Corte de Justicia de la Nación-ICON•S México, 2019, pp. 371-400.

GARGARELLA, Roberto. "Why Do We Care about Dialogue? 'Notwithstanding Clause', 'Meaningful Engagement' and Public Hearings: A Sympathetic but Critical Analysis". En: Young, Katharine G. (ed.): *The Future of Economic and Social Rights*. Globalization and Human Rights, Cambridge University Press, 2019.

GARGARELLA, Roberto. *El derecho como una conversación entre iguales. Qué hacer para que las democracias contemporáneas se abran por fin al diálogo ciudadano*. Siglo Veintiuno Editores, 2021.

GARGARELLA, Roberto. "Por una democracia ciudadana sin 'atajos'. Sobre Democracia sin atajos, de Cristina Lafont". *Revista Derecho del Estado* 55 (2023): 125-39. doi:https://doi.org/10.18601/01229893.n55.08.

GIUFFRÉ, Carlos Ignacio. "Deliberative constitutionalism 'without shortcuts': On the deliberative potential of Cristina Lafont's judicial review theory". *Global Constitutionalism* (2022): 1-19. doi:10.1017/S2045381722000211.

GIUFFRÉ, Carlos Ignacio. "Constitucionalismo Fuerte y Democracia Deliberativa: Inconsistencias en Rawls, Dworkin, y Alexy. *International Journal of Constitutional Law*, (2023a): moad071. doi:10.1093/icon/moad071.

GIUFFRÉ, Carlos Ignacio. "De la democracia deliberativa al constitucionalismo dialógico". *Revista Derecho del Estado* 55 (2023b): 141-69. doi:https://doi.org/10.18601/01229893.n55.01.

GIUFFRÉ, Carlos Ignacio. "El constitucionalismo fuerte en la encrucijada. El constitucionalismo deliberativo como salida". *Revista de Derecho Político* 118 (2023c): 289-314. doi:10.5944/rdp.118.2023.39106.

GIUFFRÉ, Carlos Ignacio. "Pushing the Boundaries of Deliberative Constitutionalism: From Judicial Dialogue to Inclusive Dialogue". *Revus* 50 (2023d). doi:10.4000/revus.9695.

GIUFFRÉ, Carlos Ignacio. "The coming of age of deliberative constitutionalism". *Canadian Journal of Law and Jurisprudence First View* (2024): 1-30. doi: https://doi.org/10.1017/cjlj.2024.14

GONZÁLEZ RICOY, Iñigo; QUERALT, Jahel. "Introducción. Tuercas y tornillos de la filosofía política". En: *Razones públicas. Una introducción a la filosofía política*. Kindle edition, Ariel, 2021.

GROBBINK, Eva; PEACH, Kathy. *Combining Crowds and Machines. Experiments in collective intelligence design 1.0*. Nesta, 2020. https://www.nesta.org.uk/report/combining-crowds-and-machines/.

HABERMAS, Jürgen. *Facticidad y validez. Sobre el derecho y el Estado democrático de derecho en términos de teoría del discurso*. Madrid: Trotta, 1998.

KONG, Hoi L.; LEVY, Ron. "Deliberative Constitutionalism". En: Bächtiger, Andre; Dryzek, John S.; Mansbridge, Jane; Warren, Mark (eds.): *The Oxford Handbook of Deliberative Democracy*. Oxford University Press, 2018.

LAFONT, Cristina. *Democracy without shortcuts. A participatory conception of deliberative democracy*. Oxford University Press, 2020.

LANDEMORE, Hélène. "Can AI bring deliberative democracy to the masses". *Working paper*, 2023. https://bit.ly/3We2LRu

LEVY, Ron. "The 'Elite Problem' in Deliberative Constitutionalism". En: Levy, Ron; Kong, Hoi L.; Orr, Graeme; King, Jeff (eds.): *The Cambridge Handbook of Deliberative Constitutionalism*. Cambridge University Press, 2018, pp. 351-69.

LEVY, Ron; KONG, Hoi L.; ORR, Graeme; KING, Jeff, eds. *The Cambridge Handbook of Deliberative Constitutionalism*. Cambridge University Press, 2018.

LINARES, Sebastián. "Democracia sin atajos: cognitivamente exigente pero participativamente austera". *Revista Derecho del Estado* 55 (2023): 57-85. doi:https://doi.org/10.18601/01229893.n55.05.

MANSBRIDGE, Jane. "Everyday talk in the deliberative system". En: Macedo, Stephen (ed.): *Deliberative Politics: Essays on Democracy and Disagreement*. Oxford University Press, 1999, pp. 211-42.

MANSBRIDGE, Jane; BOHMAN, James; CHAMBERS, Simone; CHRISTIANO, Thomas; FUNG, Archon; PARKINSON, John; THOMPSON, Dennis; WARREN, Mark. "A systemic approach to deliberative democracy". En: Mansbridge, Jane; Parkinson, John (eds.): *Deliberative Systems: Deliberative Democracy at the Large Scale*. Cambridge University Press, 2012, pp. 1-26.

MARTÍ, José Luis. *La república deliberativa. Una teoría de la democracia*. Madrid, Barcelona: Marcial Pons, 2006.

MARTÍ, José Luis. "The Role of New Technologies in Deliberative Democracy". En: Amato, Giuliano; Barbisan, Benedetta; Pinelli, Cesare (eds.): *Rule of Law vs Majoritarian Democracy*. Bloomsbury, 2021.

MARTÍ, José Luis. "Múltiples velocidades. Sobre Democracia sin atajos, de Cristina Lafont". *Revista Derecho del Estado* 55 (2023): 87-104. doi:10.18601/01229893.n55.06.

MIKHAYLOVSKAYA, Anna. "Enhancing Deliberation with Digital Democratic Innovations". *Philosophy & Technology* 37(1) (2024): 3. doi:10.1007/s13347-023-00692-x.

NOVECK, Beth Simone. *Wiki government: how technology can make government better, democracy stronger, and citizens more powerful*. Washington, D.C: Brookings Institution Press, 2009.

NOVECK, Beth Simone. "Crowdlaw: Collective Intelligence and Lawmaking". *Analyse & Kritik* 40(2) (2018): 359-80.

NOVECK, Beth Simone. "The Innovative State". *Dædalus. Journal of the American Academy of Arts & Sciences* 150(3) (2021): 121-42.

OWEN, David; SMITH, Graham. "Survey Article: Deliberation, Democracy, and the Systemic Turn". *The Journal of Political Philosophy* 23(2) (2015): 213-34. doi:https://doi.org/10.1111/jopp.12054.

PASQUALE, Frank. *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Cambridge (Mass.): The Belknap Press of Harvard University Press, 2020.

RAWLS, John. *El liberalismo político*. Barcelona: Crítica, 1996.

SHORTALL, Ruth; ITTEN, Anatol; VAN DER MEER, Michiel; MURUKANNAIAH, Pradeep; JONKER, Catholijn. "Reason against the machine? Future directions for mass online deliberation". *Frontiers in Political Science* 4 (2022): 946589. doi:10.3389/fpos.2022.946589.

SILVA, Virgílio Afonso da. "Big Brother Is Watching the Court: Effects of TV Broadcasting on Judicial Deliberation". *Verfassung Und Recht in Übersee / Law and Politics in Africa, Asia and Latin America* 51(4) (2018): 437-55.

SUSSKIND, Richard. *Online Courts and the Future of Justice*. Oxford University Press, 2019. doi:10.1093/oso/9780198838364.001.0001.

TSAI, Lily L.; PENTLAND, Alex; BRALEY, Alia; CHEN, Nuole; ENRÍQUEZ, José Ramón; REUEL, Anka. "Generative AI for Pro-Democracy Platforms". *An MIT Exploration of Generative AI* (2024). doi:10.21428/e4baedd9.5aaf489a.

VALENTINI, Chiara. "Deliberative constitutionalism and judicial review". *Revus* 47 (2018) [Online]. http://journals.openedition.org/revus/8030.

VELASCO, Juan Carlos. *Para leer a Habermas*. Madrid: Alianza Editorial, 2003.

ZURN, Christopher. *Deliberative democracy and the institutions of judicial review*. New York: Cambridge University Press, 2007.

Capítulo 14

LA INTELIGENCIA ARTIFICIAL EN EL EJERCICIO DE LA FUNCIÓN JURISDICCIONAL ¿AMENAZA O GRAN OPORTUNIDAD PARA LOS DERECHOS FUNDAMENTALES DE LOS JUSTICIABLES?

Álvaro Toribio Carrera

Doctorando en el Programa Estado de Derecho y Gobernanza Global
Universidad de Salamanca

RESUMEN: ¿Es posible sustituir al juez humano por un juez robot? ¿Podrán los sistemas de inteligencia artificial en un futuro dictar resoluciones judiciales que afecten a los derechos fundamentales y garantías procesales de los ciudadanos? Tras la crisis sanitaria consecuencia de la COVID-19 se ha acelerado el proceso de modernización y digitalización en todos los ámbitos, incluido el judicial. En ese sentido, la implementación de los instrumentos de inteligencia artificial en el ejercicio de la función jurisdiccional, capaces de imitar las capacidades cognitivas de los seres humanos para aprender y resolver problemas, origina un intenso debate jurídico y académico que obliga a los diferentes operadores jurídicos a realizar una reflexión pragmática, hermenéutica y holística del uso de esta tecnología en la toma de decisiones judiciales. Aunque la inteligencia artificial promete mejorar la eficacia y eficiencia en la toma de las decisiones judiciales, también plantea desafíos significativos en términos éticos, normativos y sociales que hacen necesario proporcionar una visión equilibrada sobre si la algoritmización judicial constituye una amenaza o una gran oportunidad para los derechos fundamentales, en especial el derecho a la tutela judicial efectiva del artículo 24.1 de la Constitución Española.

PALABRAS CLAVE: Inteligencia artificial. Sistema judicial, Juez-Robot. Derechos fundamentales. Tutela judicial efectiva. Independencia judicial. Justicia.

1. INTRODUCCIÓN

La era de la interdependencia digital [1] y la evolución súbita de las nuevas tecnologías disruptivas como la inteligencia artificial (IA), basadas en el análisis de multitud de datos y en los procesos automáticos que confunden la información con el conocimiento, obliga al humanismo tecnológico a construir una democracia que equilibre la utilización de la IA en la toma de decisiones judiciales con el uso ético de los datos e instrumentos algorítmicos que los gestionan a fin de poder advertir los potenciales riesgos, oportunidades y desafíos que los avances tecnológicos, aun en la imaginación de nuestras mentes, pronto podrían llegar a producir en la Humanidad.

La idea utópica de jueces robots autónomos fundamentando sentencias judiciales de forma rápida y eficaz mediante el procesamiento de multitud de datos y precedentes judiciales sin errores sistemáticos o sesgos humanos [2], plantea un dilema de extraordinaria importancia que excede el ámbito técnico y que aconseja ser abordado desde un punto de vista normativo, filosófico e incluso político. Aunque los avances tecnológicos nunca han ido de la mano de la Administración de Justicia y se han incorporado al proceso judicial con un evidente retraso, durante los últimos años son varios los trabajos doctrinales y académicos que abordan la posibilidad de que algún día un juez algorítmico pueda llegar a sustituir a un juez humano [3].

En cualquier caso, las dudas sobre la responsabilidad en la toma de decisiones judiciales injustas mediante el uso de la IA, la falta de imparcialidad e independencia, los sesgos discriminatorios que son inherentes a los datos de entrenamiento que procesa un algoritmo, el efecto *black box* en su diseño y funcionamiento [4] o la ausencia de empatía humana en la aplicación de la ley podrían llegar a convertir a los sistemas de IA en los verdaderos sujetos en la toma de decisiones judiciales y conculcar los derechos fundamentales de los ciudadanos, de ahí que todas estas cuestiones merezcan la atención de cualquier jurista, obligado no solo a resolver conflictos, sino a anticiparse a los problemas y ofrecer soluciones [5].

1 Informe final de El Panel de Alto Nivel sobre la Cooperación Digital, establecido por el Secretario General de las Naciones Unidas en julio de 2018.

2 FERNÁNDEZ, Carlos: "¿Se pueden evitar los sesgos en la Inteligencia Artificial?", Diario La Ley. 2021.

3 La posibilidad de aplicar la IA en los procesos judiciales no tuvo un especial interés por parte de la doctrina científica en un primer momento, si bien durante los últimos cinco años, esta ya es recurrente y pueden destacarse multitud de trabajos en la materia.

4 Libro Blanco sobre la Inteligencia Artificial: un enfoque europeo orientado a la excelencia y la confianza europea. Bruselas, 2020.

5 REBOLLO DELGADO, Lucrecio: Inteligencia artificial y derechos fundamentales, Colección IA, Robots y Bioderecho, Editorial Dykinson. 2023. pp. 14

2. SITUACION DE LA ADMINISTRACIÓN DE JUSTICIA ESPAÑOLA: DESAFIOS.

El 74 % de los españoles considera que los miembros de la Carrera Judicial son competentes y están preparados para el ejercicio de sus funciones mientras que el 82% considera que no existe posibilidad alguna de sobornar a un juez. Además, el 77% tiene una imagen de la justicia sesgada gracias a los medios de comunicación, pues solo son noticia las deficiencias del Poder Judicial pero no los miles de casos que son resueltos a diario satisfactoriamente. Ahora bien, sobre el funcionamiento de la Administración de Justicia, el 72% de los encuestados considera que la justicia es lenta y el 79% cree que carece de los recursos necesarios para actuar con rapidez y eficacia [6]. No obstante, y aunque la ciudadanía española considere como grave la eficiencia funcional de la Administración de Justicia, colapsada en algunas jurisdicciones y crónicamente desatendía de recursos personales y materiales [7], según el Consejo General del Poder Judicial en el año 2023, los órganos judiciales registraron 7.004.309 asuntos, cifra que representa un aumento del 4,8 % respecto al año 2022 [8].

Por otro lado, la histórica aversión a los cambios y el desarrollo asimétrico de la Administración de Justicia, en donde conviven la gestión analógica del papel con la gestión digitalizada, invita a incorporar en su funcionamiento diario herramientas de IA que permitan obtener una justicia moderna capaz de agilizar la resolución de los procesos judiciales, liberar de la sobrecarga de trabajo a los funcionarios de la Administración de Justicia automatizando tareas burocráticas y mejorar la interoperabilidad de los sistemas. En consecuencia, nuestro sistema judicial no puede quedar desconectado de una nueva realidad que acontece, por lo que corresponde a los poderes públicos abordar correctamente en este nuevo marco relacional, la delimitación y potenciación de este nuevo entorno digital con un propósito claro: Favorecer una más eficiente y efectiva potestad jurisdiccional garantizando los derechos fundamentales y garantías procesales de los justiciables.

3. NECESIDAD DE REGULACIÓN

El extraordinario avance de los sistemas de IA y su capacidad transformadora aconseja establecer un marco regulatorio en el ordenamiento jurídico español en el que se reconozca expresamente su uso en el ámbito de la justicia. Algunos autores hablan de la denominada "Constitucionalización del algoritmo o la digitalización de la Constitución" [9], lo que per-

6 Encuesta "Los Españoles y la Justicia" realizada por Metroscopia para el CGPJ en mayo de 2021.

7 Según el Justice Scoreboard del año 2023 de la Unión Europea, y pese de la mejora de los datos de años anteriores, la Administración de Justicia española se coloca en los puestos de cola en el tiempo para resolver los asuntos civiles y mercantiles y en proporción más baja de jueces por habitante.

8 Datos estadísticos proporcionados por el CGPJ en 2023. Disponible en: www.poderjudicial.es

9 BALAGUER CALLEJÓN, Francisco: "La Constitución del Algoritmo. El difícil encaje de la Constitución analógica en el mundo digital", en coord. Balaguer Callejón, Francisco y Cotino Hueso, Lorenzo, en *Derecho público de la inteligencia artificial*, (2023), pp.29-56.

mitiría concretar la aplicación y los límites de estos sistemas de IA en el proceso judicial sin menoscabar los principios estructurales de nuestro Estado de Derecho y en pro de una mayor eficacia y eficiencia de la Administración de Justicia.

A nivel internacional, aunque existe en la actualidad un consenso generalizado sobre la necesidad de regular la IA, lo cierto es que existe una atomización legislativa o disparidad de percepciones en la materia [10]. Sin embargo, pueden diferenciarse dos grandes modelos: De "*hard law"* y de *"soft law" [11]*. El primero de carácter regulatorio, y asumido por la mayoría de los países, se caracteriza por el establecimiento de leyes obligatorias, sanciones, procedimientos y tutela judicial mientras que el segundo se basa en recomendaciones, guías, declaraciones o códigos de conducta de carácter no vinculante [12].

3.1.Ámbito europeo

En la Unión Europea, el marco regulatorio actual es fruto de una intensa actividad política y regulatoria para dar respuesta al fenómeno de la digitalización, la implementación de la IA y el flujo de datos [13]. Centra el interés de esta comunicación, el Reglamento (UE)

10 La Universidad de Stanford en el Índice de la inteligencia artificial publicado el 3 de abril de 2023, señala que, de 127 países analizados, 31 Estados han aprobado al menos una ley de IA. A la cabeza se encuentra EE. UU. con 22 leyes, donde España ocupa el tercer puesto en el ranking mundial.

11 SALVADOR TORRES, Leopoldo y PERALTA GUTIÉRREZ, Alfonso: "Marco normativo de la IA en el ámbito comparado", en Herrera Triguero, Francisco (Coord.), Peralta Gutiérrez, Alfonso (Coord.) Torres López, Salvador (Coord.) y Artigas Brugal, Carme (pr.), *El derecho y la inteligencia artificial*, Editorial Eug, Año 2022. pp. 189-222.

12 EE. UU ha optado por priorizar la seguridad nacional obligando a las empresas tecnológicas a notificar cualquier avance que suponga un grave riesgo para la seguridad nacional, priorizando la regulación de leyes de carácter sectorial y no general de IA para incentivar las inversiones y la innovación tecnológica, aunque de forma incipiente ha comenzado a establecer recomendaciones de carácter legal a las empresas. China por su parte, ha optado por convertirse en una potencia mundial de la inteligencia artificial a través de la denominada Iniciativa de la Franja y de la Ruta mediante la firma de normas e instrumentos de cooperación con varios países. No obstante, su implementación presenta riesgos especialmente significativos, dado que no están en consonancia con el respeto a los derechos fundamentales en Occidente, ya que utilizan técnicas de vigilancia gubernamental, puntuación ciudadana o instrumentos de censura acordes a la moral y ética de la ideología de su régimen. Finalmente, en el caso de Reino Unido y como consecuencia del Brexit, esta potencia se ha centrado en garantizar el uso responsable de los datos y el desarrollo tecnológico, buscando un necesario equilibro entre la innovación y el respeto a los derechos fundamentales.

13 Entre los documentos políticos y legislativos se señalan entre otros, el Grupo de expertos de alto nivel sobre inteligencia artificial que emite un primer proyecto de directrices éticas en 2018, el Plan coordinado con los Estado miembros para el fomento y desarrollo de la IA en Europa de 2018 y actualizado en 2021,la publicación del Libro Blanco de la Inteligencia Artificial en 2020, la resolución para una propuesta legislativa de Reglamento sobre principios éticos para el desarrollo y la implementación y el uso de la inteligencia artificial la robótica y las tecnologías conexas o el Reglamento 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.

2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial, en el que la Unión Europea ha establecido la primera Ley sobre Inteligencia Artificial Europea (LIAE) que establece un modelo de IA confiable, de transparencia y explicabilidad, de minimización del sesgo de los datos y de gestión del riesgo [14] con dos objetivos: Que el ser humano y la dignidad que le es inherente sean el centro del marco normativo y que el uso de los sistemas de IA genere confianza a los ciudadanos [15].

Con respecto al ejercicio de la potestad jurisdiccional que es objeto de esta comunicación, la LIAE consciente de la afectación del uso de la IA en los derechos fundamentales, cataloga los sistemas de IA en este ámbito como de alto riesgo [16]. No obstante, deben diferenciarse dos usos de estos sistemas en el ámbito de la Administración de Justicia: Los sistemas de ayuda judicial y los sistemas de tramitación. Sobre los primeros, se advierte de la afección de estos en el derecho a la tutela judicial efectiva y en la imparcialidad e independencia judicial, por lo que deben ser considerados de alto riesgo todos aquellos sistemas que complementen a los Jueces y Magistrados a la hora de interpretar los hechos objeto de enjuiciamiento y la aplicación del ordenamiento jurídico. Respecto de los segundos, estos no afectan al ejercicio de la potestad jurisdiccional sino al funcionamiento ordinario de la Administración de Justicia a través de herramientas de anonimización o seudonimización de las sentencias judiciales, de documentos, tareas administrativas o comunicaciones judiciales, por lo que han de ser consideraros de riesgo bajo.

3.2. Ámbito nacional

En el ámbito nacional, se han establecido importantes estrategias y mecanismos de regulación en esta materia. Sirva como ejemplo, la Carta de Derechos Digitales que trata de

14 La Unión Europea a través de la LIAE ha establecido un sistema de riesgos donde se diferencia entre: 1) Sistemas de IA prohibidos, que implican un riesgo inadmisible para la seguridad, la vida y los derechos fundamentales, por ejemplo, la identificación biométrica. 2) Sistemas de IA de alto riesgo, que afectan a los derechos y libertades de las personas, aunque si bien no están prohibidos, sí están sujetos a obligaciones reforzadas que garanticen su uso legal, ético, robusto y seguro, como por ejemplo la asistencia en la interpretación jurídica y aplicación de la ley.3) Sistemas de IA de riesgo medio/bajo como los asistentes virtuales o *chatbots*. 4) Sistemas IA de riego mínimo como los filtros de spam, entre otros

15 WISNER GLUSKO, Diana Carolina: "Breves reflexiones sobre la importancia del Estado de Derecho en el desarrollo del marco legal sobre los sistemas de inteligencia artificial en la Unión Europea", Coord. GARRIDO Garrido Martin, Joaquín y Valdivia Jiménez, Ramón, en *Inteligencia artificial y filosofía del derecho,* Ediciones Laborum. 2022. pp. 529-545

16 El art 6.2 del Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, dispone que son sistemas de inteligencia artificial de alto riesgo los recogidos en el anexo III, en cuyo punto número ocho se recoge " la Administración de justicia y los procesos democráticos señalando su apartado a) los *"sistemas de IA destinados a ser utilizados por una autoridad judicial, o en su nombre, para ayudar a una autoridad judicial en la investigación e interpretación de hechos y de la ley, así como en la garantía del cumplimiento del Derecho a un conjunto concreto de hechos, o a ser utilizados de forma similar en una resolución alternativa de litigios".*

establecer la reivindicación del humanismo tecnológico analizando los principales riesgos y desafíos de la IA en los derechos fundamentales con fin de delimitar los contornos de los derechos digitales [17]. A nivel normativo, y dentro del denominado Plan Justicia de 2030 [18], se encuentran tres proyectos: De eficiencia organizativa, procesal y digital. En el programa de "Eficiencia digital" [19], se destacan tres proyectos: "La Ley de Eficiencia Digital del Servicio Público de Justicia", "la Analítica Legislativa y Judicial" y "la Seguridad Jurídica Digital", figurando entre sus subproyectos el de "Inteligencia Artificial para la eficiencia de la Justicia" [20]. Estos proyectos de ley se encontraban en tramitación parlamentaria en la anterior legislatura, pero no pudieron ser aprobados por la disolución de las Cortes en mayo de 2023, por lo que se optó por la aprobación del Real Decreto 6/2023, de 19 de diciembre, por el que se aprueban medidas urgentes para la ejecución del Plan de Recuperación, Transformación y Resiliencia en materia de servicio público de justicia, función pública, régimen local y mecenazgo que deroga la Ley 18/2011, de 5 de julio y donde se recogen muchas de las cuestiones que los proyectos de ley mencionados pretendían implementar en la Administración de Justicia. Por ejemplo, los artículos 56 a 58 del Real Decreto Ley 6/2023, de 19 de diciembre, contempla por primera vez el uso de la IA en el proceso judicial, en particular, en aquellas tareas de tramite o resoluciones judiciales simples que producen actos mecánicos o automatizados que son fruto del procesamiento de documentos y que no requieren de la interpreta-

17 Para Jiménez Rubio, la Carta de Derechos Digitales es "un documento asertivo, prospectivo y asertivo que sirve de referencia y clave de bóveda de las políticas y normas que en materia de IA se vayan aprobando a partir de ahora". En JIMENEZ RUBIO, María del Rosario: "Inteligencia artificial y registros" en Herrera Triguero, Francisco (Coord.), Peralta Gutiérrez, Alfonso (Coord.) Torres López, Salvador (Coord.) y Artigas Brugal, Carme (pr.), *El derecho y la inteligencia artificial*, Editorial Eug, Año 2022. pp. 232-256

18 Según el Ministerio de Justicia en su web oficial, Justicia 2030 "es un plan de trabajo común a 10 años, desarrollado en cogobernanza, que impulsa el Estado de Derecho y el acceso a la Justicia como palancas de la transformación de país. Sólo incide en aquellos puntos que tienen mayor impacto en el sistema o que han quedado desfasados y ya son poco operativos. No se trata de introducir cambios en cada uno de los componentes del Servicio Público de Justicia, sino de generar transformaciones en puntos que tienen efecto sistémico en el ecosistema Justicia".

19 Este programa tiene por objetivo, "la generación de un marco normativo para la digitalización en el Servicio Público de Justicia, que establezca las bases legales de la transformación digital de la Administración de Justicia y sustituya la regulación actual vigente desde 2011. La norma establecerá un nuevo marco legal para la tramitación electrónica de procedimientos judiciales, abordando aspectos básicos como la firma digital, los juicios telemáticos, el expediente judicial electrónico o el intercambio de datos en la Administración de Justicia. El objetivo es promover el Estado de Derecho en el nuevo entorno digital".

20 Siguiendo las recomendaciones de la Comisión Europea para la Eficiencia de la Justicia (CEPEJ) del Consejo de Europa, que considera que " Inteligencia Artificial puede contribuir a mejorar la eficiencia y calidad del trabajo de los tribunales" se pretende llevar a cabo según el Ministerio de Justicia el empleo de la IA para la analítica avanzada judicial, aprovechar la IA para la mejora en la gestión, explotación de información y procesos, potenciar los desarrollos legislativos orientados al dato y la regulación inteligente mediante política legislativa basadas en datos, impacto y litigiosidad asociada a las normas legales, entre otros.

ción jurídica del juez [21], así como las actuaciones proactivas [22] y asistidas, donde se genera un borrador de un documento producido por un algoritmos que puede constituir un apoyo a la resolución judicial pero que nunca constituirían una resolución judicial, ya que esta precisará de la validación del Juez o Magistrado que conozca de la causa.

Pese a estos avances regulatorios, es necesario seguir impulsando el proceso de modernización y digitalización en la justicia española para intentar dar respuesta a los verdaderos retos de organización y de fondo que subyacen en nuestro sistema judicial, lo que requiere de una regulación normativa del uso de los sistemas de IA en los textos procesales y orgánicos que regulan el ejercicio de la potestad jurisdiccional, cuyo punto de partida, además de los instrumentos citados [23], pueden ser las recomendaciones y principios reconocidos en la Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas judiciales y su Entorno de 2018, que señaló que entre los principios que han de regir la regulación de estos sistemas se encuentran el respeto a los derechos fundamentales y la prohibición de discriminación para evitar las desigualdades derivadas de los sesgos algorítmicos, la transparencia, la imparcialidad y la seguridad, así como el principio de control del usuario, gracias al cual, el justiciable goza de mecanismos de recursos y auditoria de aquellas resoluciones judiciales que se dictan usando herramientas algorítmicas.

4. LA INTELIGENCIA ARTIFICIAL JUDICIAL

Tras la invención de ENIAC a mediados del siglo XX [24], se inicia un importante pero complejo camino en el uso de computadoras en el ámbito del Derecho. Esta interacción

21 El capítulo VII del RD-Ley 6/2023 recoge los art 56 a 58 que reza "De las actuaciones automatizadas, proactivas y asistidas". En ese sentido, el art. 56 define la actuación automatizada como "aquella actuación procesal producida por un sistema de información adecuadamente programado sin necesidad de intervención humana en cada caso singular".

22 El art. 56.3 del RD-Ley 6/2023 señala que son actuaciones proactivas "aquellas actuaciones automatizadas, autoiniciadas por los sistemas de información sin intervención humana, que aprovechan la información incorporada en un expediente o procedimiento de una Administración Pública con un fin determinado, para generar avisos o efectos directos a otros fines distintos, en el mismo o en otros expedientes, de la misma o de otra Administración Pública, en todo caso conformes con la ley."

23 La Carta de Derechos Digitales alude a la necesidad de actualizar normas sectoriales que se ven afectadas por la IA, como la LO 1/1982, de 5 de mayo, de protección civil del derecho al honor, intimidad personal y familiar y propia imagen, la LO 1/1984, de 26 de marzo, reguladora del derecho de rectificación, la Ley 34/2002 de 11 de julio, de servicios de la sociedad de la información y comercio electrónico o la Ley 1072021 de 9 de julio, de trabajo a distancia entre otras. En el ámbito de la Administración de Justicia, será necesario modificar la LOPJ de 1985 y las leyes procesales, como la LECrim y LEC, a fin de regular la utilización de los instrumentos y sistemas basados en la IA para la aplicación de la ley, las herramientas de justicia predictiva y las que ayuden a los jueces y magistrados a interpretar los hechos concretos.

24 ENIAC es el acrónimo de Electronic Numerical Integrator And Computer (Computador e Integrador Numérico Electrónico) uno de los primeros computadores electrónicos de la historia, nunca imitado y el único disponible en EEUU entre 1946 y 1949. MOLERO Xavier: "ENIAC: una máquina y un tiempo por redescubrir", en Actas de las XIX Jenui. Castellón, 2013. Pp. 241-248

generará sustanciales dilemas que irán cambiando según la realidad que acontezca en cada momento histórico y evidenciará la complejidad de dicha conexión y su impacto en los derechos fundamentales de los ciudadanos.

En la actualidad, aunque el Derecho y la IA coincidan en amplias zonas de interés común, la utilización de estos sistemas en el ejercicio de la potestad jurisdiccional pondrá a prueba su intensa relación. Dado que son los seres humanos quienes han creado, desarrollado e implementado la IA, es necesario garantizar el respeto de la dignidad humana y de los derechos fundamentales consagrados en los diferentes instrumentos nacionales e internacionales de los que España forma parte, reflejando los valores humanistas en todo el proceso judicial.

4.1. Los derechos fundamentales, la IA y el ejercicio de la potestad jurisdiccional.

El sistema judicial español se encuentra condenado a la digitalización y la implementación de la IA, so riesgo de estancamiento, obsolescencia y petrificación. No obstante, la falta de control y transparencia sobre el algoritmo capaz de complementar o adoptar decisiones judiciales, puede impactar en los derechos fundamentales de los ciudadanos, especialmente en el derecho a la tutela judicial efectiva, el derecho de defensa, la imparcialidad o independencia judicial o el principio de publicidad de las actuaciones judiciales.

En una primera aproximación, y de conformidad con lo establecido en el art 117. 3 de la Constitución Española, los Jueces y Magistrados integrados en el Poder Judicial, ejercen la potestad jurisdiccional en todo tipo de procesos, juzgando y haciendo ejecutar lo juzgado, cuya legitimación democrática deriva del pueblo y de su sometimiento con exclusividad al imperio de la ley. A tenor del señalado precepto, ser juzgado por un juez humano independiente e imparcial se configura no solo como un derecho fundamental de nuestra Constitución, tal y como proclama nuestra Carta Magna, sino como un requisito inherente a cualquier sociedad democrática que pretenda garantizar la paz y el orden social. Téngase en cuenta que, en nuestro ordenamiento jurídico, la emisión de una sentencia judicial fundada, congruente y acorde al derecho aplicable por parte de los titulares de los diferentes órganos jurisdiccionales que conocen de un asunto, constituye el culmen de un proceso jurídico riguroso que tiene como objetivo hacer justicia y proteger los derechos fundamentales de todas las partes involucradas en el proceso judicial. De acuerdo con esto, el ejercicio de la función jurisdiccional y la emisión de una sentencia no es un simple acto formal, sino el resultado de un análisis jurídico minucioso llevado a cabo por un juez que debe garantizar la congruencia fáctica y jurídica y que estará irremediablemente ligada a la idea de justicia. Así pues, el juez humano no solo es un servidor público que aplica la ley de forma automatizada y estandarizada, sino que es capaz de interpretar el dinámico y cambiante ordenamiento jurídico y buscar soluciones conforme a Derecho dentro de las posibilidades que el propio ordenamiento jurídico le brinda, sin caer en el puro arbitrio.

4.2. ¿Es posible un juez-robot? La IA como complemento de decisiones judiciales.

El debate académico, jurídico y ético sobre la sustitución del juez humano por el juez IA presenta multitud de matices y aristas. Téngase en cuenta además que el término "Inteligencia Artificial" es un oxímoron literario, dado que ninguna maquina tiene conciencia de sí misma ni la capacidad de inteligencia intuitiva o emocional de un juez humano, pues únicamente están sujeta a la dictadura de los datos que procesa un algoritmo.

Sobre la idea de sustituir al juez humano por un juez robot, se perfilan dos posiciones doctrinales perfectamente diferenciadas: En contra y a favor. Los detractores descartan la posibilidad del juez robot al esgrimir la necesaria reforma constitucional del artículo 117.1 y 3 de la Constitución, la imposibilidad actual de la tecnología existente para motivar resoluciones judiciales o advierten que, prescindir del factor humano en la toma de decisiones judiciales es inviable dado que la aplicación de la ley al caso concreto requiere de un margen de maniobra o discrecionalidad técnica que requiere necesariamente del factor humano. Frente a los negacionistas, los partidarios del uso de la IA en la toma de decisiones judiciales consideran que nada impide que el vertiginoso avance de la tecnología y su perfeccionamiento permita que los sistemas de IA puedan resolver los asuntos concretos con cierta discrecionalidad al igual que hacen los jueces humanos, siempre que así lo establezca el algoritmo. Además, los sistemas de IA podrían supervisar las decisiones judiciales a fin de garantizar que se adoptan soluciones similares ante supuestos análogos, reduciendo la discrecionalidad técnica del juez humano o incluso podrían detectar los sesgos derivados de la naturaleza humana al tomar decisiones, lo que potenciaría la imparcialidad judicial ya que los sistemas de IA podrían detectar los sesgos cognitivos o heurísticos epistemológicos del juez humano gracias al concurso y complemento de la IA en la toma de decisiones judiciales [25].

Sea como fuere, el objetivo de esta comunicación es determinar si actualmente la figura del juez robot podría realizar el ejercicio de la función jurisdiccional sin el menoscabo de los derechos constitucionales recogidos en el artículo 24 de la Constitución Española y las garantías procesales de cualquier ciudadano sometido a un proceso judicial. Es decir, si los jueces robots podrían proporcionar una motivación jurídica acorde al ordenamiento jurídico vigente sobre unos hechos fácticos controvertidos [26].

25 *"…los jueces no operan en el vacío. Ni pueden ni deben hacer lo. De hecho, ningún ser humano lo hace en su actuar cotidiano. Lo que sucede en el exterior nos condiciona, nos aflige y llega a afectar nuestra toma de decisiones... el juzgador no pue de actuar aislado del entorno social y las decisiones que este toma en los litigios de interés público son fruto de un diálogo natural, permanente e inevitable con otros elementos sociales y políticos —académicos, prensa, el poder ejecutivo, el legislador, agencias administrativas y el público, en general".* En SIMÓN CASTELLANO, P: "Justicia cautelar e inteligencia artificial. La alternativa a los atávicos heurísticos judiciales", J. M. Bosch Editor, Barcelona, 2021, pp. 260 y ss.

26 Desde un punto de vista comparativo, en países de nuestro entorno como Reino Unido, Francia o Estonia, se han iniciado proyectos que actualmente están en suspenso para implementar en el ámbito civil y penal, el establecimiento de sistemas predictivos y de toma de decisiones judiciales automatizadas ante supuestos sencillos y bajo la concurrencia de determinadas circunstancias. En el resto de los casos más complejos y no estandarizados, las decisiones adoptadas por la IA servirían como instrumento de apoyo al juez que, ante las opciones

A este respecto, son varios los autores que contemplan el establecimiento de sistemas inteligentes capaces de adoptar decisiones judiciales ante la concurrencia de determinadas circunstancias, si bien limitan esta posibilidad a procesos que carezcan de complejidad, tenga un contenido análogo, exista uniformidad jurisprudencial sobre la cuestión y la presencia de prueba documental. Según este posicionamiento, la presencia de jueces robots resultaría factible en procedimientos de cláusulas abusivas, juicios rápidos, procesos por decreto, procedimientos monitorios sin oposición o condiciones generales de contratación, entre otros [27]. En consecuencia, los sistemas de IA con algoritmos transparentes y explicables serían capaces de adoptar decisiones judiciales automatizadas en determinadas áreas jurisdiccionales, siempre que se trate de ámbitos perfectamente acotados con jurisprudencia estable y consolidada, si bien estas transformaciones deberían ir acompañadas de la promoción y garantía de los principios esenciales del proceso judicial. Otros defensores de los sistemas de IA en la toma de decisiones judiciales auguran que en las próximas décadas *"los sistemas de IA superarán a los jueces dictando sentencias razonadas"* [28], señalando que aunque los algoritmos de los sistemas de IA también presentan sesgos en su diseño, no debe olvidarse que estos derivan en última instancia de los sesgos inherentes a los seres humanos que se incorporan a los datos que el propio algoritmo analiza. Así, los jueces humanos estarían influenciados por circunstancias internas y externas [29], incluso por el propio sistema que los lleva en ocasiones a dictar soluciones distintas ante supuestos análogos, a diferencia de los sistemas de IA, capaces de adoptar la misma decisión siempre que los datos proporcionados sean los mismos.

Por su parte, los detractores de esta tecnología en la toma de decisiones judiciales afirman que los sistemas de IA serian incapaces de emular la inteligencia humana, ya que carecen de elementos cognitivos y volitivos capaces de aportar los matices que solo un juez humano puede proporcionar debido a su naturaleza. En ese sentido, la IA sería incapaz de combinar los elementos jurídicos y fácticos de una disputa jurídica, usar analogías para dar respuestas a antinomias existentes en el ordenamiento jurídico, fijar la verdadera esencia de las controversias jurídicas o recurrir a valores y principios no escritos que les permitiría ir más allá del análisis objetivo y estadísticos de los datos y las consecuencias lógicas y practicas derivados de los mismos. Solo la labor de un juez humano podría, tras analizar las circunstancias concurrentes, la legislación vigente aplicable al caso y el estudio de la jurisprudencia previa, comprender el fondo del asunto y discernir, conforme a su experiencia y pericia, una solución

propuestas por el sistema de IA y que no serían vinculantes, tendría que motivar en su argumentación jurídica la elección o descarte de la propuesta ofrecida por el sistema de IA en su sentencia.

27 DELGADO MARTÍN, Joaquín: "Judicial-Tech, el proceso digital y la transformación tecnológica de la Justicia". Madrid, La Ley.2020.

28 Susskind ha publicado desde hace más de 30 años, innumerables obras en materia de IA y sistema judicial, tales como El futuro de la ley (1996), Transformando la ley (2000), ¿El fin de los abogados? (2008), Tribunales en línea y el futuro de la justicia (2019).

29 El profesor Nieva, alude a los denominados heurísticos o sesgos cognitivos como ideas preconcebidas, creencias, tradiciones sociales y jurídicas, forma de gestionar las emociones, personalidad del juez que pueden conducir a automatismos en la toma de decisiones judiciales. En NIEVA FENOLL, Jordi: "Inteligencia artificial y proceso judicial". Madrid: Marcial Pons. 2018

justa mientras que los sistemas de IA solo podrían complementar la labor judicial, liberando al juez de trabajos automatizados y estandarizados como la traducción de textos, el análisis de datos o la síntesis de información para centrar todo su talento y capacidad jurídica en la búsqueda de soluciones justas alejadas de la objetividad de silogismos que únicamente llegan a soluciones necesarias e inexorables fruto de patrones que no garantizan la idea de justicia.

En otro orden de cosas, el uso de algoritmos puede impactar negativamente en la imparcialidad y garantías procesales que han de observarse a lo largo de cualquier procedimiento judicial, ya sea porque los algoritmos presentan sesgos desde su creación y diseño o bien porque su uso quedase al margen del control y supervisión humano. No obstante, estos riesgos no impiden que, establecidas las cautelas necesarias, los sistemas de IA puedan sean utilizados para apoyar y complementar la función jurisdiccional siempre que se permita el acceso a su diseño y contenido a los miembros de la Carrera Judicial asi como a los justiciables.

En definitiva, los sistemas de IA son esclavos de su propio algoritmo, buscando patrones que se repiten y que les impide hacer algo que su algoritmo les prohíbe, por lo que en el ámbito judicial es deseable que el uso de los instrumentos de IA esté necesariamente ligado únicamente a apoyar y complementar la toma de decisiones judiciales, previa supervisión y evaluación ex ante y ex post por parte de los Jueces y Magistrados que conozcan de un determinado asunto. A tal efecto, no resulta aconsejable sustituir los jueces humanos por jueces robots, a fin de evitar caer en la denominada "dictadura digital"[30] donde la IA puede contener sesgos de diseño y funcionamiento que pueden dar resultados desviados no conocidos por el juez que ha de dictar una resolución judicial [31].

5. CONCLUSIONES

La disrupción tecnológica de la IA en el sistema judicial está transformando la forma de entender la justicia, rediseñando el proceso de resolución de conflictos. La heterogeneidad

30 Nieva Fenoll señala que estaríamos pasando de la justicia de los jueces a la justicia de los programadores y de aquéllos que les influyan, lo cual es un riesgo inasumible en cualquier Estado democrático y de Derecho. En NIEVA FENOLL, Jordi: "Inteligencia artificial y proceso judicial". Madrid: Marcial Pons. 2018

31 En el ámbito europeo, la normativa opta por no permitir a los sistemas de IA llevar a cabo la toma de decisiones judiciales sin el necesario control y supervisión humana. El juez no quedaría en ningún caso vinculado por el borrador propuesto por el sistema de IA, pudiendo modificar o adoptar una decisión distinta a la planteada por los sistemas de IA que permitan la producción de actuaciones judiciales y procesales automatizadas, asistidas y proactivas. En España, y pese a lo establecido en el art 56 a 58 del Real Decreto 6/2023, de 19 de diciembre que contempla el uso de la IA en el proceso judicial, en particular, en aquellas tareas de tramite o resoluciones judiciales simples que producen actos mecánicos o automatizados que no requieren de la interpretación jurídica del juez, el uso de la IA en la toma de decisiones judiciales no plantearía problemas. No obstante, en aquellas decisiones que requieren de interpretación jurídica, el "juez IA", no se contempla, pues sería preceptiva además de una reforma constitucional a tenor de lo establecido en el art 117.1 y 3 de la CE, la inclusión de tecnologías de argumentación jurídica capaces de incorporar a los pronunciamientos de los jueces robots una motivación que cumpliera con las exigencias derivadas del artículo 24 de la CE.

de las posibilidades de su uso, la complejidad y la evolución del propio concepto de lo que es la IA obliga necesariamente a los juristas a repensar y madurar las bases fácticas, éticas y normativas existentes de los sistemas de IA antes de positivizarlas en instrumentos normativos con vocación temporal de permanencia. Frente al entusiasmo cauteloso de los partidarios de la IA en el sistema judicial y la sospecha paciente de sus detractores, el uso de esta tecnología debe de ir acompañada de mayores exigencias públicas y del desarrollo de nuevos principios que permitan garantizar la exactitud, explicabilidad, trazabilidad y publicidad de estos sistemas, así como la posibilidad de recurrir las decisiones adoptadas por ellos. Del mismo modo, es perentoria la cooperación, colaboración y coordinación entre la Administración de Justicia, el mundo académico y judicial, la industria y los organismos públicos para la creación de órganos independientes encargados de verificar la objetividad de los algoritmos con el objetivo de evitar sesgos y prevenir los efectos discriminatorios de los mismos.

Por otro lado, la IA constituye una gran oportunidad para garantizar el derecho a la tutela judicial efectiva consagrado en el art 24.1 de la Constitución Española. Gracias a estos sistemas, se pueden analizar multitud de datos a gran escala, detectar patrones y generar posibles predicciones, lo que reforzaría la labor de Jueces y Magistrados a la hora de adoptar sus decisiones judiciales, ya que estos contarían con todos los instrumentos normativos y precedentes jurisprudenciales para motivar su resolución judicial. Además de ganar tiempo y mejorar la confiabilidad del desempeño de su labor jurisdiccional, los miembros de la Carrera Judicial podrían enfocar su capacidad en tareas argumentativas más complejas que requieren obligatoriamente del componente cognitivo humano. Del mismo modo, estos sistemas vanguardistas permitirían minimizar los riesgos asociados de su implementación, reforzando la seguridad y confianza de los ciudadanos en el sistema judicial donde su correcta adecuación potenciaría la independencia e imparcialidad judicial reduciendo el margen de apreciación subjetivo derivado de la propia naturaleza humana. No obstante, consideramos que una decisión judicial que fuese fruto únicamente de la decisión automatizada de un algoritmo proveniente de un sistema de IA, seria contraria al derecho a la tutela judicial efectiva y las garantías procesales por contravenir el derecho a la motivación de las resoluciones judiciales del articulo 120.3 de la Constitución y por impedir la acción judicial efectiva contra la decisión adoptada por los sistemas de IA.

Finalmente, se ha de destacar que la labor creativa de los jueces a la hora de adoptar sus decisiones judiciales mediante la calificación de los hechos, la valoración de las pruebas, la ponderación de los intereses en conflicto y la determinación de la norma aplicable, combina una serie de elementos de discrecionalidad técnica, contextualización y motivación judicial que permiten detectar y resolver falacias, lagunas y antinomias garantizando la idea de justicia. Esta idea es inalcanzable para los jueces robots, ya que solo serían capaces de aportar una solución correcta pero no necesariamente justa. Además, no debe olvidarse que ninguna maquina podría ser totalmente autónoma y libre, dado que carece de dignidad, elemento diferenciador del juez humano frente el juez robot. Por consiguiente, el juez humano difícilmente será reemplazado por los sistemas de IA si los miembros de la Carrera Judicial son capaces de sobreponerse al avance tecnológico y evitan que el ejercicio de la función jurisdiccional quede ligado a los intereses de empresas privadas titulares de dicha tecnología, pues no debe

olvidarse que la función de juzgar no consiste en buscar soluciones únicamente repetitivas y legales sino justas, transformadoras y preventivas que permitan mejorar la vida en sociedad.

6. BIBLIOGRAFÍA

BALAGUER CALLEJÓN, Francisco: "La Constitución del Algoritmo. El difícil encaje de la Constitución analógica en el mundo digital", en coord. Balaguer Callejón, Francisco y Cotino Hueso, Lorenzo, en *Derecho público de la inteligencia artificial*, (2023), pp.29-56.

BARONA VILAR, Silvia: "Algoritmización del Derecho y de la Justicia. De la inteligencia artificial a la Smart Justice." Valencia: Tirant lo Blanch. 2021.

COTINO HUESO, Lorenzo: "Riesgos e impactos del big data, la inteligencia artificial y la robótica. Enfoques, modelos y principios de la respuesta del derecho", en Revista General de Derecho Administrativo, 50. 2017.

DELGADO MARTÍN, Joaquín: "Judicial-Tech, el proceso digital y la transformación tecnológica de la Justicia". Madrid, La Ley.2020.

DE HOYOS SANCHO, Montserrat: "El uso jurisdiccional de los sistemas de inteligencia artificial y la necesidad de su armonización en el contexto de la Unión Europea", en Revista General de Derecho Procesal, nº 55. 2021.

FERNÁNDEZ, Carlos: "¿Se pueden evitar los sesgos en la Inteligencia Artificial?", Diario La Ley. 2021.

GÓMEZ COLOMER, Juan Luis: "El Juez Robot. La independencia judicial en peligro." Tirant lo Blanch, Valencia, España. 2023.

JIMENEZ RUBIO, María del Rosario: "Inteligencia artificial y registros" en Herrera Triguero, Francisco (Coord.), Peralta Gutiérrez, Alfonso (Coord.) Torres López, Salvador (Coord.) y Artigas Brugal, Carme (pr.), *El derecho y la inteligencia artificial*, Editorial Eug, Año 2022. pp. 232-256

LASALLE RUIZ, José María: "Humanismo tecnológico y Ciberleviatán: una respuesta cívica a la distopía digital", en Francisco Herrera Triguero (coord.), Alfonso Peralta Gutiérrez (coord.), Leopoldo Salvador Torres López (coord.), Carme Artigas Brugal (pr.): El derecho y la Inteligencia Artificial, España: Editorial Eug. 2022. pp. 33-49.

MARTIN DIZ, Fernando: "Inteligencia artificial y proceso: Garantías frente a eficiencia en el entorno de los derechos procesales fundamentales", en Jiménez Conde, Fernando (dir.), Bellido Penadés, Rafael (dir.), Llopis Nadal, Patricia y Luis García, Elena (coords.), en Justicia: ¿garantías "versus" eficiencia?, Tirant lo Blanch. 2019.

NIEVA FENOLL, Jordi: "Inteligencia artificial y proceso judicial." Madrid: Marcial Pons. 2018.

REBOLLO DELGADO, Lucrecio: "Inteligencia artificial y derechos fundamentales." Colección IA, Robots y Bioderecho, Editorial Dykinson. 2023.

ROBERTO GRANERO, Horacio: "Derechos y garantías concretas frente al uso de inteligencia artificial y decisiones automatizadas, especialmente en el ámbito judicial y de aplicación de la ley", en Lorenzo Cotino Hueso (dir.), Marcelo Bauzá Reilly (coord.), en Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas, Aranzadi. 2022. pp. 107-137.

SALVADOR TORRES, Leopoldo y PERALTA GUTIÉRREZ, Alfonso: "Marco normativo de la IA en el ámbito comparado", en Herrera Triguero, Francisco (Coord.), Peralta Gutiérrez, Alfonso (Coord.) Torres López, Salvador (Coord.) y Artigas Brugal, Carme (pr.), *El derecho y la inteligencia artificial*, Editorial Eug, Año 2022. pp. 189-222.

SANCHIS CRESPO, Carolina: "Inteligencia artificial y decisiones judiciales: crónica de una transformación anunciada", SCIRE: Representación y organización del conocimiento (2023) Vol. 29, n.º 2, pp. 65-84.

SOLAR CAYÓN, José Ignacio: "¿Jueces-robot? Bases para una reflexión realista sobre la aplicación de la inteligencia artificial en la Administración de Justicia", en Solar Cayón, José Ignacio y Sánchez Martínez, Mª Olga (dir.), en El impacto de la inteligencia artificial en la teoría y la práctica jurídica, La Ley (Wolters Kluwer), Madrid. 2022. pp. 245-280.

SUSSKIND, Richard: "Tribunales on-line y la Justicia del futuro." La Ley-Wolters Kluwer. 2020.

WISNER GLUSKO, Diana Carolina: "Breves reflexiones sobre la importancia del Estado de Derecho en el desarrollo del marco legal sobre los sistemas de inteligencia artificial en la Unión Europea", Coord. GARRIDO Garrido Martin, Joaquín y Valdivia Jiménez, Ramón, en *Inteligencia artificial y filosofía del derecho,* Ediciones Laborum. 2022. pp. 529-545

CAPÍTULO 15

MAPPING THE INFLUENCE OF ARTIFICIAL INTELLIGENCE: A PROPOSAL FOR EFFECTIVE VISUALIZATION IN BUSINESS AND PUBLIC ADMINISTRATION

Antoni MESTRE
Researcher at the Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València

Victoria TORRES
Researcher at the Valencian Research Institute for Artificial Intelligence and professor at Universitat Politècnica de València

ABSTRACT: The integration of artificial intelligence (AI) into business and public administration sectors has introduced both transformative opportunities and significant challenges. As AI technologies become increasingly embedded in various organizational and governmental processes, understanding their impact requires sophisticated tools for effective visualization. This paper addresses the need for such tools by proposing the Sustainability Awareness Diagram (SusAD), a framework designed to offer a structured approach to visualize the multifaceted effects of AI. The SusAD framework aims to simplify the complex nature of AI systems, enhance decision-making, and improve communication among stakeholders by translating intricate AI processes and impacts into intuitive visual formats. Through this visualization tool, stakeholders can better identify potential risks, opportunities, and interdependencies associated with AI integration. The paper concludes by highlighting the importance of further research to refine visualization tools, validate their effectiveness through empirical studies, and explore additional metrics for a comprehensive understanding of AI impacts. Future research should also focus on integrating visualization tools with other decision-support systems and engaging diverse stakeholders to tailor tools to their needs. By advancing these areas, the study aims to support the ethical and sustainable deployment of AI technologies in business and public administration.

KEYWORDS: AI Governance · AI Visualization · AI Regulation · AI Impact Analysis · Ethical AI

1. INTRODUCTION

The advent of artificial intelligence (AI) has initiated a paradigm shift in various sectors, notably in the corporate world and public administration (Samoili et al. 2021). The capabilities of AI to process vast amounts of data, learn from patterns, and make autonomous decisions have revolutionized traditional workflows and service delivery methods. These advancements promise significant improvements in efficiency, accuracy, and personalization of services, positioning AI as a cornerstone of modern innovation. However, alongside these benefits, the integration of AI technologies also brings forth complex challenges, particularly concerning the ethical, social, and economic impacts, as well as legal and policy implications (Bankins and Formosa 2023).

In the corporate environment, AI technologies are being deployed to automate routine tasks, optimize supply chains, enhance customer service through chatbots, and facilitate data-driven decision-making processes (Almgren and Skobelev 2020). For instance, AI-driven analytics can uncover insights from big data that were previously unattainable, enabling companies to better understand market trends and consumer behavior. This, in turn, leads to more informed strategic decisions and competitive advantages. Similarly, in public administration, AI applications are streamlining operations, improving the efficiency of public services, and enabling more personalized interactions with citizens (Henman 2020). Examples include automated processing of administrative tasks, predictive analytics for public health management, and AI-powered tools for urban planning and resource allocation (Herath and Mittal 2022).

Despite these promising developments, the rapid and widespread adoption of AI raises critical concerns (Hernández-Orallo 2017). Ethical considerations, such as the potential for bias in AI algorithms, the impact on employment due to automation, and overarching issues of privacy and data security, are at the forefront of public and academic discourse (Kazim and Koshiyama 2021). Additionally, the opaque nature of many AI systems, often described as "black boxes," complicates efforts to understand and mitigate potential adverse effects. This lack of transparency can hinder the ability of stakeholders to assess the broader implications of AI integration effectively.

Legal and policy considerations also play a significant role in shaping the deployment of AI technologies. The need for robust regulatory frameworks to ensure fairness, accountability, and transparency in AI systems is imperative (Rodrigues 2020). Policymakers are tasked with balancing the innovation-driven benefits of AI with the protection of public interests, such as safeguarding privacy rights, preventing discrimination, and ensuring equitable access to technology. These regulatory frameworks must evolve in tandem with technological advancements to address emerging risks and challenges (Smuha 2021).

In response to these challenges, there is a pressing need for tools that can facilitate a comprehensive and nuanced understanding of the impacts associated with AI technologies (Chatzimparmpas et al. 2020). One such tool is the Sustainability Awareness Diagram (SusAD) (Penzenstadler et al. 2018). The SusAD is designed to visualize the chains of effects

resulting from the integration of new technologies, such as AI. It serves both as a compilation tool and a facilitation mechanism for discussions among stakeholders, including requirements engineers, system designers, policymakers, and end-users.

The SusAD operates by capturing potential effects in a structured diagram, prompting stakeholders to reflect on how one effect may lead to another over time and across various dimensions—social, economic, ethical, environmental, and legal. This approach not only aids in identifying immediate impacts but also encourages consideration of long-term consequences and interdependencies. By providing a clear and intuitive visualization of these chains of effects, the SusAD supports more informed decision-making and policy development, fostering an environment where the benefits of AI can be maximized while mitigating potential risks.

This paper aims to explore the necessity and utility of the SusAD in the context of AI integration in both the workplace and public administration. We will discuss the current landscape of AI applications, highlight the associated challenges, and demonstrate how the SusAD can be employed to address these issues. Through a practical example, we illustrate the effectiveness of the SusAD in facilitating comprehensive evaluations and informed discussions among diverse stakeholders.

Overall, the integration of AI technologies presents both significant opportunities and profound challenges. As AI continues to evolve and permeate various aspects of society, the need for tools like the SusAD becomes increasingly critical. By providing an intuitive and comprehensive means to visualize and discuss the impacts of AI, the SusAD can play a pivotal role in ensuring that these technologies are deployed ethically, effectively, and sustainably.

2. SUSTAINABILITY AND THE NEED FOR VISUALIZATION

AI has become a transformative force across various sectors, including business and public administration. Its capabilities to process vast amounts of data, learn from patterns, and make autonomous decisions have led to significant advancements and efficiencies. However, the integration of AI also presents complex challenges that require a deeper understanding and effective management (Wirtz, Weyerer and Sturm 2020). This section discusses the importance of sustainability in AI and the necessity of visualization tools to address the intricate impacts of AI technologies.

2.1. The Sustainable Artificial Intelligence

AI technologies encompass a broad range of applications, from machine learning and natural language processing to robotics and autonomous systems. In business, AI enhances operational efficiency through automation, predictive analytics, and customer relationship manage-

ment. For instance, AI-driven algorithms can optimize supply chain logistics, tailor marketing strategies, and streamline customer service operations. In public administration, AI holds promise for improving service delivery, public safety, and policy implementation (Henman 2020). Predictive policing, automated administrative tasks, and smart city initiatives are examples of how AI can augment governmental functions and enhance citizen engagement.

Despite these benefits, the deployment of AI systems introduces several challenges. The opaque nature of AI decision-making processes, often referred to as the "black box" problem, complicates the understanding of how decisions are made. This lack of transparency can lead to issues such as biased outcomes, accountability concerns, and ethical dilemmas. As AI systems become more integral to decision-making processes, it becomes crucial to develop mechanisms that can elucidate their workings and impacts.

AI and other emerging technologies must adhere to principles of sustainability to ensure their benefits do not come at the cost of long-term environmental and societal health. The concept of sustainability, as defined by the Oxford English Dictionary (Simpson et al. 1989), is "the capacity to endure." The Brundtland Commission further elaborates on sustainable development as "meeting the needs of the present without compromising the ability of future generations to meet their needs" (Development 1987). In the context of AI, sustainability involves ensuring that technological advancements contribute positively across various dimensions, including individual, social, economic, technical, and environmental.

Given the complex nature of AI systems and their potential impacts, there is a pressing need for robust visualization tools to map out and manage these effects. Effective visualization tools are crucial for simplifying the intricate networks of cause and effect inherent in AI deployments. They help stakeholders—ranging from policymakers to technical experts and the general public—understand the potential outcomes and implications of integrating AI into various sectors.

2.2. The Need for Visualization

Given the complexity and potential risks associated with AI technologies, there is a pressing need for effective visualization tools. Visualization serves as a powerful means to represent complex data and processes in an accessible and interpretable manner. For AI systems, visualization tools can bridge the gap between technical intricacies and stakeholder understanding, facilitating better decision-making and communication (Beauxis-Aussalet et al. 2021).

2.2.1. Enhancing Understanding

AI systems often operate with intricate algorithms and large datasets that can be challenging to grasp in their raw form. Visualization tools help in simplifying these complexities

by providing clear and intuitive representations of how AI technologies function and influence various aspects of business and public administration. For example, visualizing the flow of data through an AI model can help stakeholders understand how inputs are transformed into outputs and how different variables interact within the system.

2.2.2. Supporting Decision-Making

Effective visualization tools enable stakeholders to gain valuable insights into the potential outcomes of AI integration. By presenting data in visual formats, such as charts, graphs, and diagrams, stakeholders can more easily assess the implications of different AI strategies and make informed decisions. Visualization supports the evaluation of various scenarios, helping to balance the potential benefits of AI with its associated risks and challenges.

2.2.3. Improving Communication and Transparency

One of the primary benefits of visualization is its ability to enhance communication and transparency. AI systems often involve multiple stakeholders, including developers, policymakers, and the public. Visualization tools facilitate clearer communication by translating complex AI processes and impacts into formats that are understandable to non-technical audiences. This improved transparency helps build trust and fosters more collaborative discussions about AI implementation and governance.

2.2.4. Identifying Risks and Opportunities

Visualization tools play a crucial role in identifying potential risks and opportunities associated with AI technologies. By mapping out the interactions and dependencies between various components of an AI system, stakeholders can anticipate possible challenges and uncover areas for optimization. For instance, visualizing the potential impacts of an AI-driven policy can reveal unintended consequences and help devise strategies to mitigate risks.

3. PROPOSING AN ESTABLISHED VISUALIZATION FRAMEWORK

In addressing the multifaceted impacts of AI in business and public administration, a robust visualization framework is essential. Visualization tools help to simplify complex data and processes, making it easier for stakeholders to understand and manage the effects of AI

technologies. This section introduces the proposed framework—the SusAD—and outlines its components, methodology for implementation, and benefits.

SusAD is a comprehensive visualization tool designed to capture and represent the multifaceted impacts of AI technologies. It provides a structured approach to mapping out the effects of AI systems across various dimensions, including operational, social, ethical, and environmental aspects. The SusAD framework is built on the premise that effective visualization can enhance understanding, facilitate informed decision-making, and improve stakeholder communication.

3.1. Components of the SusAD

The SusAD consists of two primary components: sustainability dimensions and temporal levels of impact. These components work together to provide a comprehensive framework for understanding and managing the effects of AI technologies.

3.1.1. Sustainability Dimensions

The SusAD is structured as an adapted radar chart divided into five equal parts, each representing one of the sustainability dimensions: individual, social, economic, technical, and environmental (Penzenstadler and Femmer 2013).

Individual Dimension: This dimension considers the impact of technologies on personal freedom, agency, human dignity, and fulfillment. It includes the ability of individuals to thrive, exercise their rights, and develop freely. Technologies can enhance personal experiences and quality of life but may also pose risks to privacy, autonomy, and mental well-being.

For AI, this dimension examines how AI systems influence individual autonomy, privacy, and personal development. For example, AI-driven personalized recommendations can enhance user satisfaction, but invasive data collection practices might compromise privacy and individual agency.

Social Dimension: This dimension examines the relationships between individuals and groups, focusing on mutual trust, communication, and the balance of conflicting interests within a social system. Technologies can facilitate social interactions and enhance community building, but they can also lead to social fragmentation, digital divides, and erosion of trust.

With AI, this dimension explores the impact of AI systems on social dynamics and community structures. AI-powered social media platforms can connect people globally, fostering community bonds, but they can also contribute to misinformation, polarization, and social isolation.

Economic Dimension: This dimension looks at the financial aspects and business value generated by technologies, including capital growth, liquidity, investment, and financial

operations. Technologies can drive economic growth and create new business opportunities, but they can also disrupt existing industries and exacerbate economic inequalities.

In AI, this dimension assesses how AI technologies affect economic activities and business models. AI-driven automation can boost productivity and economic efficiency, but might also lead to job losses and economic disparity if not managed properly.

Technical Dimension: This dimension addresses the maintenance and evolution of artificial systems over time, emphasizing resilience, maintenance, and ease of system transitions. Technologies should be designed for longevity, adaptability, and reliability, considering technical challenges and risks.

For AI, this dimension evaluates the sustainability of AI systems in terms of their technical robustness, scalability, and adaptability. AI systems need to be resilient to failures, capable of evolving with new data, and easy to maintain to ensure long-term functionality and reliability.

Environmental Dimension: This dimension evaluates the use and stewardship of natural resources, covering immediate waste production, energy consumption, local ecosystems, and climate change concerns. Technologies can help reduce environmental impact through efficiency and innovation, but they can also lead to increased resource consumption and environmental degradation if not managed sustainably.

Regarding AI, this dimension looks at the environmental footprint of AI technologies. AI development and deployment consume significant computational resources, leading to high energy usage and associated carbon emissions. Sustainable AI practices aim to minimize this impact through energy-efficient algorithms and environmentally conscious data center management.

3.1.2. Temporal Levels of Impact

The three temporal levels represented in the SusAD—immediate, enabling, and structural effects—help visualize the short-term, medium-term, and long-term impacts of technologies. This temporal distinction is crucial for understanding how effects evolve and interact over time.

Immediate Effects: These are direct functions of the system or direct effects of its development, observable shortly after implementation. Immediate effects of AI include improvements in efficiency and decision-making in various applications, such as automated customer service and real-time data analysis.

Enabling Effects: These arise from the use of the system and may become apparent over a medium timeframe, influencing broader applications and user behaviors. Enabling effects of AI might include increased accessibility to services through AI-powered tools and enhanced capabilities in predictive analytics that support proactive decision-making.

Structural Effects: These refer to persistent changes that can be observed at the macro-level over a long period, often altering fundamental aspects of society and environment.

Structural effects of AI encompass long-term shifts in job markets due to automation, changes in societal norms and behaviors influenced by AI-driven personalization, and significant environmental impacts from sustained computational demands.

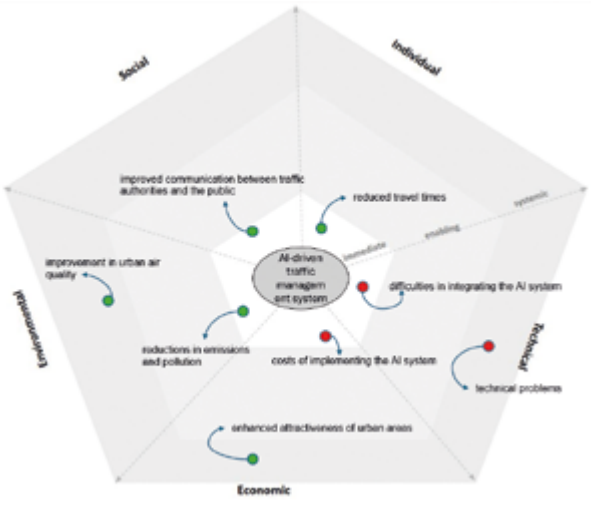
By integrating these sustainability dimensions and temporal levels of impact, the SusAD provides a comprehensive framework for assessing and visualizing the sustainability impacts of AI technologies. This holistic approach ensures that stakeholders can anticipate and manage the diverse and evolving effects of AI in a balanced and informed manner.

3.2. Example of SusAD Application

To demonstrate the practical use of the SusAD, we present a case study on the implementation of an AI-driven urban traffic management system. This example shows how the SusAD can visualize and manage the sustainability impacts of AI technologies throughout their lifecycle.

Urban traffic congestion is a major issue in many cities, causing longer travel times, higher fuel consumption, and increased greenhouse gas emissions. To address these problems, city planners propose an AI-driven traffic management system aimed at optimizing traffic flow, reducing congestion, and minimizing environmental impact.

SusAD is employed to evaluate the necessity and suitability of the AI system for urban traffic management. This initial phase involves workshops with a diverse group of stakeholders, including city planners, environmental scientists, local business owners, and residents. The SusAD framework facilitates a thorough exploration of the AI system's potential impacts across five sustainability dimensions.



Individual Dimension: The SusAD helps assess the immediate effects of the AI system on individuals, such as reduced travel times and decreased stress for commuters.

Social Dimension: In terms of social impacts, the SusAD identifies immediate benefits such as improved communication between traffic authorities and the public through real-time traffic updates.

Economic Dimension: The SusAD framework evaluates the immediate economic effects, such as the upfront costs associated with implementing the AI system. It also considers structural economic impacts, including the long-term benefits derived from increased efficiency and the enhanced attractiveness of urban areas due to better traffic management.

Technical Dimension: The feasibility study uses the SusAD to identify immediate technical challenges, such as difficulties in integrating the AI system with existing infrastructure. It also addresses potential structural issues, such as technical problems or the obsolescence of older systems that may arise as the AI system is integrated.

Environmental Dimension: Finally, the SusAD assesses the immediate environmental effects, including reductions in emissions and pollution due to optimized traffic flow. It also evaluates long-term environmental impacts, such as improvements in urban air quality and public health resulting from sustained reductions in pollution.

By applying the SusAD in the feasibility study phase, stakeholders can gain a comprehensive understanding of the AI system's potential impacts across multiple dimensions of sustainability, facilitating informed decision-making and strategic planning for the project.

3.3. Benefits of Using the SusAD

The SusAD offers a comprehensive framework for visualizing and managing the impacts of AI technologies across multiple dimensions of sustainability. Its implementation provides several significant benefits, which are crucial for both researchers and practitioners in the field of AI and sustainability.

3.3.1. Holistic Impact Assessment

One of the primary benefits of using the SusAD is its ability to provide a holistic view of the impacts of AI technologies. By incorporating multiple sustainability dimensions—individual, social, economic, technical, and environmental—the SusAD ensures that all relevant factors are considered. This comprehensive approach helps identify and address potential negative impacts early in the technology's lifecycle, facilitating more balanced and sustainable outcomes.

3.3.2. Enhanced Stakeholder Engagement

The SusAD serves as an effective communication tool that facilitates stakeholder engagement and collaboration. Its visual nature allows for easier interpretation and discussion of complex data and relationships. By involving stakeholders from the initial study phase through to post-implementation, the SusAD helps ensure that their concerns and insights are integrated into the decision-making process. This inclusive approach not only enhances transparency but also builds trust and buy-in among all parties involved.

3.3.3. Improved Decision-Making

The SusAD aids in making more informed and strategic decisions by clearly illustrating the cause-and-effect relationships of AI technology impacts over time. The three temporal levels—immediate, enabling, and structural effects—help decision-makers understand both short-term and long-term consequences. This temporal perspective is critical for anticipating potential risks and opportunities, enabling proactive management and continuous improvement.

3.3.4. Facilitation of Sustainability Goals

By explicitly mapping out the impacts of AI technologies across different sustainability dimensions, the SusAD aligns the technology implementation process with broader sustainability goals. It provides a structured framework for identifying how AI can contribute to or detract from these goals. This alignment is particularly important for organizations committed to sustainable development, as it helps ensure that AI technologies are deployed in ways that support long-term environmental, social, and economic well-being.

3.3.5. Versatility and Adaptability

The SusAD is a versatile tool that can be adapted to various contexts and technologies beyond AI. Its flexibility allows it to be used in different phases of technology implementation, from initial feasibility studies to long-term impact assessments. This adaptability makes the SusAD a valuable asset for a wide range of applications, ensuring its relevance and utility across different projects and sectors.

3.3.6. Enhanced Monitoring and Reporting

The ongoing use of the SusAD during the implementation and post-implementation phases provides a robust mechanism for monitoring and reporting. By regularly updating the diagram with new data, organizations can track the progress of AI technology impacts in real-time. This continuous monitoring helps identify emerging issues, measure the effectiveness of mitigation strategies, and document success stories. The ability to generate detailed, visual reports enhances accountability and supports continuous improvement efforts.

3.3.7. Promotion of Interdisciplinary Research

The SusAD encourages interdisciplinary research by integrating diverse perspectives and expertise. Its multi-dimensional approach necessitates collaboration between different fields, such as computer science, environmental science, economics, and social sciences. This interdisciplinary nature fosters innovative solutions and comprehensive analyses, contributing to the advancement of knowledge and practice in sustainable technology implementation.

4. CONCLUSIONS AND FUTURE RESEARCH

The integration of AI into business and public administration sectors has unveiled both significant potential and complex challenges. This paper has underscored the necessity for effective visualization tools to comprehend and manage the impact of AI technologies. By introducing the SusAD as a proposed framework, this study aims to address these needs by offering a structured approach to visualize the multifaceted effects of AI.

One of the key conclusions drawn from this research is the critical role of visualization in simplifying the complex nature of AI systems and their impacts. The SusAD framework offers a clear and intuitive means to map out how AI technologies influence various dimensions of business and public administration. This visualization facilitates a better understanding of AI's effects, enabling stakeholders to gain valuable insights into the potential outcomes of AI integration. As a result, it supports more informed decision-making processes, helping to balance the benefits of AI with its associated risks.

Another significant conclusion is that visualization improves communication and transparency by bridging gaps between technical and non-technical stakeholders. By translating complex AI processes and impacts into understandable visual formats, visualization enhances communication and fosters more collaborative discussions. Additionally, the SusAD framework helps in identifying potential risks and opportunities related to AI integration. By visualizing interdependencies and potential effects, organizations can proactively address challenges and leverage advantages.

Looking forward, several areas warrant further research to advance the understanding and application of AI visualization. There is a need for ongoing refinement and enhancement of visualization tools such as SusAD. Future research should focus on developing more sophisticated methodologies for capturing and representing the complex interactions and effects of AI systems. Empirical validation is also crucial; conducting real-world case studies across different industries and public sectors will provide insights into the practical applications of the SusAD framework and identify potential improvements.

Furthermore, exploring additional metrics and dimensions for inclusion in visualization frameworks will contribute to a more comprehensive understanding of AI impacts. This exploration should delve deeper into social, ethical, and environmental aspects. Additionally, future research should examine how visualization tools can be integrated with other decision-support systems and frameworks, including risk assessment models, compliance monitoring systems, and strategic planning tools.

Engaging diverse stakeholders in the development and use of visualization tools is also vital. Research should focus on understanding the needs and perspectives of various user groups to better tailor tools to their requirements. Lastly, investigating the long-term impacts of AI technologies using visualization tools is important for understanding how AI evolves over time and its subsequent effects on organizational and societal outcomes.

By addressing these research areas, future studies can contribute to more effective and comprehensive approaches for managing AI technologies, ultimately supporting their ethical and sustainable deployment.

5. REFERENCES

ALMGREN, R. and SKOBELEV, D., 2020. Evolution of Technology and Technology Governance. *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 6, no. 2, ISSN 2199-8531. DOI 10.3390/joitmc6020022.

BANKINS, S. and FORMOSA, P., 2023. The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics*, vol. 185, no. 4, ISSN 1573-0697. DOI 10.1007/s10551-023-05339-7.

BEAUXIS-AUSSALET, E., BEHRISCH, M., BORGO, R., CHAU, D.H., COLLINS, C., EBERT, D., EL-ASSADY, M., ENDERT, A., KEIM, D.A., KOHLHAMMER, J., OELKE, D., PELTONEN, J., RIVEIRO, M., SCHRECK, T., STROBELT, H. and VAN WIJK, J.J., 2021. The Role of Interactive Visualization in Fostering Trust in AI. *IEEE Computer Graphics and Applications*, vol. 41, no. 6, ISSN 1558-1756. DOI 10.1109/MCG.2021.3107875.

CHATZIMPARMPAS, A., MARTINS, R.M., JUSUFI, I., KUCHER, K., ROSSI, F. and KERREN, A., 2020. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum*, vol. 39, no. 3, ISSN 1467-8659. DOI 10.1111/cgf.14034.

DEVELOPMENT, W.C. on E. and and DEVELOPMENT, W.C. on E. and, 1987. *Our Common Future*. Oxford, New York: Oxford University Press. ISBN 978-0-19-282080-8.

HENMAN, P., 2020. Improving public services using artificial intelligence: possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration*, vol. 42, no. 4, ISSN 2327-6665. DOI 10.1080/23276665.2020.1816188.

HERATH, H.M.K.K.M.B. and MITTAL, M., 2022. Adoption of artificial intelligence in smart cities: A comprehensive review. *International Journal of Information Management Data Insights*, vol. 2, no. 1, ISSN 2667-0968. DOI 10.1016/j.jjimei.2022.100076.

HERNÁNDEZ-ORALLO, J., 2017. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. S.l.: Cambridge University Press. ISBN 978-1-316-94320-5.

KAZIM, E. and KOSHIYAMA, A.S., 2021. A high-level overview of AI ethics. *Patterns* [en línea], vol. 2, no. 9, [consulta: 24 julio 2024]. ISSN 2666-3899. DOI 10.1016/j.patter.2021.100314. Disponible en: https://www.cell.com/patterns/abstract/S2666-3899(21)00157-4.

PENZENSTADLER, B., DUBOC, L., VENTERS, C.C., BETZ, S., SEYFF, N., WNUK, K., CHITCHYAN, R., EASTERBROOK, S.M. and BECKER, C., 2018. Software Engineering for Sustainability: Find the Leverage Points! *IEEE Software*, vol. 35, no. 4, ISSN 1937-4194. DOI 10.1109/MS.2018.110154908.

PENZENSTADLER, B. and FEMMER, H., 2013. A generic model for sustainability with process- and product-specific instances. *Proceedings of the 2013 workshop on Green in/by software engineering* [en línea]. New York, NY, USA: Association for Computing Machinery, pp. 3-8. [consulta: 24 julio 2024]. GIBSE '13, ISBN 978-1-4503-1866-2. DOI 10.1145/2451605.2451609. Disponible en: https://doi.org/10.1145/2451605.2451609.

RODRIGUES, R., 2020. Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*, vol. 4, ISSN 2666-6596. DOI 10.1016/j.jrt.2020.100005.

SAMOILI, S., LOPEZ-COBO, M., DELIPETREV, B., PLUMED, F., GÓMEZ, E. and DE PRATO, G., 2021. *AI Watch. Defining Artificial Intelligence 2.0. Towards an operational definition and taxonomy of AI for the AI landscape*. S.l.: s.n.

SIMPSON, J., WEINER, E., SIMPSON, J. and WEINER, E., 1989. *The Oxford English Dictionary*. Second Edition, Second Edition. Oxford, New York: Oxford University Press. ISBN 978-0-19-861186-8.

SMUHA, N.A., 2021. From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology*, vol. 13, no. 1, ISSN 1757-9961. DOI 10.1080/17579961.2021.1898300.

WIRTZ, B.W., WEYERER, J.C. and STURM, B.J., 2020. The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration*, vol. 43, no. 9, ISSN 0190-0692. DOI 10.1080/01900692.2020.1749851.

Capítulo 16

HACIA LA CONSTRUCCIÓN DE ESTÁNDARES MÍNIMOS DE PROTECCIÓN DE LOS DERECHOS DE LOS CONTRIBUYENTES EN EL PROCEDIMIENTO LEGISLATIVO TRIBUTARIO ASISTIDO POR INTELIGENCIA ARTIFICIAL[1]

Carlos E. Weffe H.
Profesor Ayudante Doctor
de Derecho Financiero y Tributario
Universitat Jaume I

RESUMEN: El artículo propone cinco estándares mínimos con el fin de regular el uso de IA en el proceso legislativo tributario. La implementación de estas reglas básicas reduciría significativamente los riesgos que supone el uso indiscriminado de IA para la adecuada salvaguarda de los derechos fundamentales de los contribuyentes: (i) a la reserva legal tributaria, (ii) a la participación ciudadana en el proceso legislativo y (iii) a la libertad económica de los contribuyentes, en especial en sus facetas de ingreso y permanencia en sus mercados relevantes, que tiene como corolarios la libertad negocial y el derecho a la economía de opción en materia tributaria, y con ello (iv) la tributación conforme a la capacidad económica. En la formulación de estos estándares mínimos se ha tenido especialmente en cuenta los principios de IA propuestos por la OCDE de IA centrada en el ser humano, respeto por los derechos humanos e implementación inclusiva e imparcial, además de la «reserva de humanidad» (PONCE SOLÉ) evitación de sesgos, calidad de datos (*garbage in, garbage out*) y supervisión humana (*human in the loop*). La investigación propuesta complementa la identificación de estándares mínimos y mejores prácticas para la protección de los derechos de los contribuyentes de BAKER y PISTONE (2015), bajo una metodología deductiva en la línea de la adaptación de los principios clásicos de protección de los derechos de los contribuyentes intentada por BENTLEY (2019).

PALABRAS CLAVE: Inteligencia artificial, procedimiento legislativo, reserva de humanidad, derechos de los contribuyentes, estándares mínimos.

1 Esta investigación fue completada durante la estancia del autor como Investigador Postdoctoral en Fiscalidad e Inteligencia Artificial de la Universitat Oberta de Catalunya, bajo la dirección de la Prof. Dra. Ana María Delgado García, a quien el autor desea agradecer cálidamente.

1. INTRODUCCIÓN

La IA es hoy, antes que un *desiderátum*, un dato de la realidad. Todas los individuos e instituciones se sirven, cada vez más, de las herramientas que componen el «*universo IA*»: sistemas generativos, discriminativos o explicativos, reactivos, de memoria limitada, de aprendizaje con distintos niveles de supervisión[2] o por refuerzo, procesadores de lenguaje natural, «visión» por ordenador,[3] sistemas «expertos»,[4] sistemas de recomendación y robótica; basados en reglas, redes neuronales y aprendizaje profundo o evolutivo (RUSSELL y NORWIG, 2010).

Naturalmente, la organización estatal no escapa a esta tendencia. En nuestra disciplina, el uso de IA en las administraciones tributarias (AATT) es -de hecho- cada vez más relevante. Según el Foro de Administración Tributaria de la OCDE, "muchas administraciones se han embarcado en un viaje de transformación digital" (OCDE, 2021). Por ejemplo, en mayo de 2024 la Agencia Estatal de Administración Tributaria (AEAT) ha apostado de manera "decidida" por el uso de IA como tecnología esencial para la mejora de la eficiencia en la consecución de sus objetivos, particularmente en lo relativo a la asistencia al contribuyente y en la prevención y lucha contra el fraude fiscal y aduanero (AEAT, 2024).

Los avances en la capacidad de los sistemas de cómputo basados en redes neuronales para «aprendizaje» (*machine learning*), el aumento de los datos disponibles y el desarrollo de nuevos algoritmos han catalizado la tendencia alcista. Para 2022 el 53% de las Administraciones Tributarias de la OCDE utilizaban modelos predictivos basados en IA para anticipar riesgos de incumplimiento y tomar las medidas respectivas (OCDE, 2022).

En efecto, las posibilidades de optimización de los procesos de organizaciones públicas son infinitas, abarcando no sólo la actividad administrativa, sino todas las formas de la actividad estatal en materia tributaria. En un apretado resumen, estas posibilidades pueden enumerarse del siguiente modo:

1.1. La IA en la política fiscal

La política fiscal, entendida como el conjunto de medidas tendientes al diseño de un sistema tributario capaz de sufragar el nivel necesario de gasto público en la forma más eficiente e igualitaria posible (TANZI Y SEE, 2000), es potenciada por la IA a través de la identificación de los patrones «anormales» de conducta, potencialmente tendientes al incumplimiento de las obligaciones tributarias. Sobre la base de esta *alerta temprana*, es posible la modificación

2 En función del nivel de «etiquetado» de los datos previos, para facilitar su clasificación y el hallazgo de patrones.

3 El que sirve de base a sistemas como los de reconocimiento facial o de diagnóstico médico por imágenes.

4 Particularmente interesantes para nuestra disciplina, en la medida en las que su especificidad les permite dar soluciones concretas y justificadas a problemas en un campo acotado, de modo similar al del juicio humano.

normativa «teledirigida» para reducir las lagunas regulativas que permiten la erosión de bases imponibles, el traslado de beneficios tributarios y, en resumen, la elusión y la evasión fiscal (WEFFE, 2021).

1.2. La IA en la gestión tributaria

Es relativamente común el uso de *chatbots* para prestar asistencia al contribuyente. Los asistentes virtuales se emplean en Australia, Brasil, España, Estados Unidos, Finlandia, Guatemala, México, Perú y Reino Unido, entre otros (COLLOSA, 2024). Adicionalmente, la IA se emplea para la formulación de declaraciones electrónicas prellenadas y el análisis comparativo de declaraciones previas para su comprobación. En 2023, la *Australian Taxation Office* (ATO) usó análisis de datos por IA para cumplimentar previamente las declaraciones de la renta, cruzar datos de declaraciones tributarias y requerir de los contribuyentes la comprobación de sus declaraciones, en los casos en que la declaración previa difiere de la de otros contribuyentes en circunstancias similares (IBFD OBSERVATORY ON THE PROTECTION OF TAXPAYERS' RIGHTS, 2024).

1.3. La IA en la inspección tributaria

Ya se mencionó el uso de IA como base de los sistemas de alerta temprana para la detección de tendencias anormales y la selección de contribuyentes. Es conocido el caso de HERMES, una herramienta desarrollada por la AEAT para análisis de riesgos y la asignación de modalidades de intervención (OLIVER CUELLO, 2021). La experiencia comparada brinda muchos ejemplos del uso de modelos de riesgo para identificar posibles diferencias de tributo y prevenir el fraude «carrusel» en el IVA: Australia, Brasil, Canadá, Estados Unidos, Francia, India, Malta, Italia, Suecia, Países Bajos, Polonia, Rumanía, Singapur y Vietnam (ASQUITH, 2024). El análisis de datos se ha extendido a la calificación y cuantificación de los hechos imponibles para la liquidación del tributo. En India, la AT ha efectuado varias liquidaciones por deducciones indebidas, con apoyo de IA en 2023 (SEN, 2023). Otro tanto ha ocurrido en Polonia y Estados Unidos (SUTTON, 2023). En España, la doctrina ha tratado el caso jurisprudencial en el que el sistema «Renta Web» calculó el importe de las amortizaciones deducibles en el IRPF, en perjuicio de los derechos a la buena administración, a la motivación de los actos y a la tutela efectiva (DE HARO IZQUIERDO, 2021).

1.4. La IA en la recaudación tributaria

En este ámbito, varias AATT realizan predicción de insolvencias para priorizar la recaudación ejecutiva, como por ejemplo Finlandia, Irlanda, Singapur y Suecia (GARCÍA-HE-

RRERA BLANCO, 2020). La comunicación con los contribuyentes insolventes a través de *chatbots* también ha favorecido la recaudación: para 2023, estas herramientas han intercambiado más de 1,6 millones de mensajes con dichos contribuyentes (SUTTON, 2024). El análisis de datos a través de IA ha soportado también la detección de incrementos patrimoniales no justificados en Costa Rica, España, Estados Unidos, México y Reino Unido, entre otros (COLLOSA, 2021).

1.5. La IA en la resolución de controversias fiscales

Hoy se acepta casi sin reparos que la IA sirva como apoyo para la realización de tareas concretas en materia de gestión procesal (RUIBAL PEREIRA, 2021). Ahora, si bien el avance en otras disciplinas jurídicas es mayor -como en la justicia penal- (STATE V. LOOMIS, 2017) (TAYLOR, 2023), la intervención de las TIC en el juzgamiento de casos tributarios ha estado siempre entre los objetivos de la Jurimétrica (LOEVINGER, 1971), el Análisis Económico del Derecho (WEFFE, 2023(a)) y la Informática Judicial (SUSSKIND, 1986). Recuérdese que uno de los primeros clásicos de la literatura en *AI & Law* es *Taxman*, un experimento de IA diseñado para «aplicar» la legislación fiscal estadounidense para «resolver» casos sobre fiscalidad de reorganizaciones empresariales (MCCARTY, 1977). *Taxman* no es sino uno de muchos sistemas «expertos» en fiscalidad, cuyo crecimiento ha sido exponencial desde entonces (SURDEN, 2019) (GÓMEZ REQUENA, 2022).

Tal explosión en el uso de la IA en el sector público en general (TORRECILLA SALINAS, 2023) ha despertado las alarmas naturales respecto de la protección de los derechos fundamentales. Se plantea -una vez más- un problema de racionalidad axiológica, que contrapone la eficacia de la actividad estatal con la indiscutible preeminencia de la dignidad humana como fundamento del ordenamiento y como límite insalvable de la acción del Estado.

La intersección entre la IA y los derechos de los contribuyentes ha generado copiosa bibliografía, que ha arribado a conclusiones de provecho aplicables al uso de IA por las AATT. Sin perjuicio de que esa labor aún esté incompleta, esto sugiere una aproximación al problema distinta: esto es, reflexionar sobre la mejor forma *práctica* de protección de los derechos de los contribuyentes frente al uso de IA por el Estado *en general*.

Este documento pretende identificar estándares mínimos de uso de IA en el *proceso legislativo*, en continuación de la reflexión del autor -y de otros académicos- (RANGONE, 2023) sobre el punto. De este modo, se pretende realizar una contribución adicional (WEFFE, 2023 (b)) en el estudio del uso de IA en la legislación, un fenómeno *insular* donde "los académicos apenas se citan entre sí, lo que sugiere que rara vez leen el trabajo de los demás" (RANGONE, 2023).

Aquí conviene aclarar que, aunque la similitud con el planteamiento de BAKER y PISTONE (2015) es evidente, el objetivo de estas líneas no es *inducir* mejores prácticas con base

en datos de la experiencia comparada,[5] sino *deducir* estándares mínimos de protección del contribuyente (WEFFE, 2017) con base -de momento- en una aproximación teórica, más cercana -probablemente- a la adaptación intentada por BENTLEY (2019).

Ello no obsta para que se constate la adecuación de los estándares ya identificados por BAKER y PISTONE para la protección de los derechos de los contribuyentes frente al uso de IA en el proceso legislativo. En adelanto de lo que se expondrá enseguida, es necesario puntualizar que las propuestas contenidas en este documento *complementan*, no sustituyen, el estándar mínimo de limitación taxativa de los casos de retroactividad de leyes tributarias y las mejores prácticas de (i) prohibición de retroactividad y (ii) participación ciudadana en el proceso legislativo a través de consulta pública previa que BAKER Y PISTONE propusieron en 2015.

Para lograr ese objetivo, la sección 2 identificará algunas de las principales áreas de conflicto entre el uso de la IA en el proceso legislativo y los derechos de los contribuyentes. Sobre esta base, la sección 3 formulará los estándares mínimos que den respuesta a estas cuestiones, y evaluará brevísimamente su pertinencia, dada la fuerte limitación de espacio a la que este documento está sometido. Estos estándares deben garantizar la protección de los derechos de los contribuyentes y ser consistentes con los principios constitucionales. Además, deben promover la transparencia, la equidad, la responsabilidad y la seguridad en el uso de IA.

En última instancia, se pretende proporcionar recomendaciones prácticas para la implementación de IA en el ámbito legislativo tributario, garantizando un equilibrio adecuado entre la eficiencia administrativa y la protección de los derechos de los contribuyentes.

2. IA, LEGISLACIÓN TRIBUTARIA Y DERECHOS FUNDAMENTALES. CIERTAS ÁREAS DE CONFLICTO

Algunos de los riesgos que supone el empleo de IA en el proceso de formación legislativa tributaria, especialmente bajo el llamado *AI oriented approach*, han sido identificadas por el autor en otra ocasión. Esta sección resumirá esos hallazgos (WEFFE, 2023 (b)).

En primer lugar, la reserva legal y la legalidad tributaria están sometidas al riesgo del llamado «legalismo computacional» acuñado por DIVER (2022): una especie extrema de *fetichismo legal*, de normas que, por ser *redactadas* por IA, están aparentemente dotadas de la certidumbre que da la ciencia y que, por ello, pueden (deben) ser *irreflexivamente* aceptadas por la ciudadanía.

En segundo lugar, por esa misma razón es posible que una adopción acrítica del *AI oriented approach* en la preparación de leyes tributarias: (i) obstaculice la participación plena del electorado en la elaboración normativa con independencia del *notice and comment*, al permitir la categorización de la relevancia de los comentarios reproduciendo sesgos; y (ii) haga

5 Tal ejercicio se deja para una ocasión posterior.

irrelevante al legislador por el descenso de su calidad o incluso por su sustitución, vaciando de contenido al *consentimiento informado* de la representación política que fundamenta la reserva legal tributaria (RANGONE, 2023).

Una *tercera* fuente de riesgo proviene de la utilización de IA para la identificación de patrones «anormales» de conducta de los contribuyentes como fundamento de la modificación normativa tributaria «teledirigida». En exceso de su función de reducir lagunas regulativas para reducir o eliminar la elusión y la evasión fiscal, el uso de IA para la tipificación de hechos imponibles y caracterización de bases imponibles puede conducir a la eliminación del derecho del contribuyente a la economía de opción y -por esa vía- crear una barrera de entrada de los agentes económicos a sus mercados relevantes, con los consecuentes daños al *status* jurídico de los contribuyentes (WEFFE, 2021) (WEFFE, 2023 (b)).

3. ESTÁNDARES MÍNIMOS DE PROTECCIÓN DE LOS DERECHOS DE LOS CONTRIBUYENTES EN EL PROCESO LEGISLATIVO APOYADO POR IA: UNA PROPUESTA

Esta sección propone cinco estándares mínimos cuya implementación, a juicio del autor, reduciría significativamente los riesgos que el uso indiscriminado de IA -sobre todo, se insiste, sobre la base del llamado *AI oriented approach*- supone para la adecuada salvaguarda de los derechos fundamentales de los contribuyentes. Se busca proteger los derechos de los contribuyentes: (i) a la reserva legal tributaria -en especial, a la *reserva parlamentaria*- (GARCÍA DE ENTERRÍA Y FERNÁNDEZ, 1983), (ii) a la participación ciudadana en el proceso legislativo y (iii) a la libertad económica de los contribuyentes, en especial en sus facetas de ingreso y permanencia en sus mercados relevantes, que tiene como corolarios la libertad negocial y el derecho a la economía de opción en materia tributaria -y con ello, claro está, (iv) la tributación conforme a la capacidad económica.

3.1. El empleo de IA en el proceso legislativo (*rule modeling*) debe estar limitado a proveer de información relevante al redactor humano (*information oriented approach*)

En la discusión sobre IA y sector público son ya tópicas dos afirmaciones: una, que en general la IA debe asistir -no sustituye- al ser humano; y dos, que los procesos de decisión automatizada deben estar supervisados -previa o posteriormente- por un humano (PÉREZ BERNABÉU, 2022), independientemente de la participación de seres humanos en el entrenamiento algorítmico, el llamado «*human in the loop*» (MOSQUEIRA-REY *et al.*, 2023).

Dadas (i) la naturaleza *ablatoria* de la prestación tributaria, en sí misma limitativa de derechos fundamentales (ANDRADE RODRÍGUEZ, 2018) y, en especial, del derecho de propiedad; y (ii) el amplio margen de configuración del legislador tributario, parece necesario

defender la necesidad de imponer un límite cualitativo a la toma de decisiones discrecionales automatizadas *falsamente habilitadas* por el legislador que supone la adopción del *AI oriented approach*.

Como enseña PONCE SOLÉ (2019), aunque él lo hace en el terreno de la relación jurídico-administrativa, la IA podrá hacerse servir como apoyo, pero la ponderación final que conduzca a la decisión debe ser humana. En otras palabras, también de PONCE SOLÉ, es necesario establecer una reserva para la toma de decisiones relativas al diseño técnico del tributo y su consagración legislativa a humanos: la «reserva de humanidad».

3.2. En todo caso, el Parlamento debe controlar que los proyectos de leyes tributarias preparados con apoyo de IA generativa no reproduzcan los sesgos que puedan deducirse de los datos que «instruyen» al algoritmo (*garbage in, garbage out, human in the loop*)

Si bien lo señalado anteriormente no obsta -como apunta VOERMANS (S/F)- a que el *AI oriented approach* no pueda ser útil para crear herramientas informáticas para *partes específicas* de la redacción legislativa o sistemas de apoyo a la toma de decisiones para la aplicación de la legislación, parece de Perogrullo afirmar que es necesario el control de los datos que alimentan el proceso de *machine learning* algorítmico, de modo que la IA generativa no reproduzca los eventuales *sesgos* que los datos puedan contener. Como con acierto indica GIL GARCÍA (2022), "por muy bien que pueda haberse diseñado una herramienta o técnica de IA, si los datos que se han utilizado en su creación son de baja calidad el resultado alcanzado puede ser deficiente".

Al hilo del estándar mínimo propuesto anteriormente, la participación humana en el control del entrenamiento de la IA generativa utilizada en el contexto legislativo («*human in the loop*») es, así, fundamental.

3.3. Todas las decisiones relevantes en las fases de redacción y discusión de leyes tributarias deben realizarse y controlarse por seres humanos (reserva de humanidad)

Como corolario -o como base, según se vea- de lo dicho anteriormente, el autor cree evidente la necesidad de sostener que en el proceso legislativo tributario es imperativo aplicar irrestrictamente la reserva de humanidad de la que habla PONCE SOLÉ (2019). A ello conducen, como se ha desarrollado previamente, el carácter ablatorio del poder tributario y el amplísimo margen de configuración de los tributos reconocido al legislador (TEDH, 2013).

3.4. El uso de IA para automatizar procesos de consulta pública de leyes tributarias debe garantizar la participación plena del cuerpo político en la elaboración normativa, evitando sesgos

Como se indicó previamente, el empleo indiscriminado de IA generativa bajo los parámetros del *AI oriented approach* puede llevar al llamado «legalismo computacional». Este modelo puede conducir, bajo la «ilusión de objetividad» de la que habla HILDEBRANDT (2020), a la adopción acrítica de nueva legislación, sin permitir -o limitando indebidamente- la participación ciudadana en el proceso legislativo, a través de la consulta pública. Ello, además, con independencia de que el contenido así aprobado sea inconstitucional por violación de la garantía de la igualdad ante la ley.

3.5. El contribuyente debe tener derecho a la impugnación de la constitucionalidad del producto legislativo, sobre la base del error en la "completitud lógica" del *rule as code*

Por definición, todo uso estatal de la IA debe ser impugnable (KAMINSKI Y URBAN, 2021), máxime cuando de ello depende la protección de varios de los pilares fundamentales sobre los que se basa el sistema tributario y, con él, el sistema democrático. No puede olvidarse que entre los mayores riesgos derivados del uso de IA en el proceso legislativo está la propia representatividad democrática del Parlamento, que en nuestra disciplina se expresa a través del apotegma *no taxation without repesentation*.

4. CONCLUSIONES

El uso de la IA en el proceso legislativo tributario, si bien supone mejoras importantes en el tratamiento y análisis de los datos relevantes para la formulación normativa de, implica también riesgos de calado para derechos y garantías fundamentales de los contribuyentes como los de reserva legal, legalidad, participación ciudadana, libertad económica, economía de opción y capacidad económica.

En este sentido, la implementación de IA para la adopción de legislación tributaria requiere de reglas básicas, estándares mínimos, que permitan la integración práctica de los principios relativos a la IA de la OCDE (2019) *basados en valores*, en especial los relativos a: (i) IA centrada en el ser humano; (ii) transparencia e inteligibilidad; (iii) derechos humanos y valores democráticos, equidad y privacidad; (iv) explicabilidad; y (v) responsabilidad, así como de los derechos fundamentales.

Bajo este contexto, es posible deducir una serie de estándares mínimos que deben regir en la práctica la actuación de los órganos legislativos al asistirse de IA. Así las cosas, a juicio

del autor el empleo de IA en el proceso legislativo (i) debe limitarse a proveer de información relevante al redactor humano (*information oriented approach*); (ii) todas las decisiones relevantes en la redacción y discusión de leyes tributarias deben realizarse por seres humanos (reserva de humanidad); (iii) debe garantizar la calidad de los datos a través del control humano, para evitar sesgos (*garbage in, garbage out* y *human in the loop*); (iv) debe garantizar la participación plena del cuerpo político en la elaboración normativa, evitando sesgos; y (v) debe ser susceptible de impugnación, tanto en su uso como en sus resultados, sobre la base -entre otras- de la «completitud lógica» del *rule as code* como fórmula de *rule modeling*.

5. BIBLIOGRAFÍA

AGENCIA ESTATAL DE ADMINISTRACIÓN TRIBUTARIA: *Estrategia de Inteligencia Artificial*. Madrid, AEAT, 2024, p. 3. Disponible en: https://rb.gy/22v8bw. Consultado el 19/7/2024.

ANDRADE RODRÍGUEZ, Betty: Inmunidad *Tributaria de los Derechos Humanos, Capacidad Contributiva y Mínimo Vital*. Ciudad de México, Thomson Reuters Dofiscal, 2018.

ASQUITH, Richard: *New models of Artificial Intelligence (AI) and Machine Learning (ML) are being evaluated by global tax authorities to tackle fraud*. VATCalc (2024), sitio web. Disponible en: https://bit.ly/4d9hzXg. Consultado el 19/7/2024.

BAKER, Philip y PISTONE, Pasquale: "General Report" en *The Practical Protection of Taxpayers' Fundamental Rights*. Basilea, IFA Cahiers vol. 100B, 2015, pp. 17-99.

BENTLEY, Duncan: "Timeless Principles of Taxpayer Protection: How They Adapt to Digital Disruption". *eJournal of Tax Research* (2019), 18(3), pp. 679-713. Disponible en: https://bit.ly/4d9KquF. Consultado el 20/7/2024.

COLLOSA, Alfredo: *Asistentes conversacionales virtuales en las Administraciones Tributarias: El futuro es hoy*. Centro Interamericano de Administraciones Tributarias (2024), Panamá, sitio web. Disponible en: https://bit.ly/3Y9wjkD. Consultado el 19/7/2024.

COLLOSA, Alfredo: *Uso de Big Data en las Administraciones Tributarias*. Centro Interamericano de Administraciones Tributarias (2021), Panamá, sitio web. Disponible en: https://bit.ly/3WvcQtM. Consultado el 19/7/2024.

DE HARO IZQUIERDO, Miguel: "Robótica, Inteligencia Artificial y Derechos Constitucionales en los Procedimientos Tributarios de Comprobación", en SERRANO ANTÓN, Fernando (Dir.) *Inteligencia Artificial y Administración Tributaria: Eficiencia Administrativa y Defensa de los Derechos de los Contribuyentes.* Cizur Menor, Thomson Reuters Aranzadi, 2021, pp. 233-253.

DIVER, Laurence E.; *Digisprudence. Code as Law Rebooted.* Edinburgo, Edinburgh University Press, 2022, p. 10. Disponible en: https://bit.ly/3Q2jOC6. Consultado el 13/10/2023.

GARCÍA DE ENTERRÍA, Eduardo y FERNÁNDEZ, Tomás-Ramón: *Curso de Derecho Administrativo*. Tomo I. Cuarta edición. Madrid, Editorial Civitas, Madrid, 1983, pp. 241-242.

GARCÍA-HERRERA BLANCO, Cristina: *El uso de la Inteligencia Artificial por las Administraciones fiscales, una cuestión de principios*. Centro Interamericano de Administraciones Tributarias (2020), Panamá, sitio web. Disponible en: https://bit.ly/3zXSpfT. Consultado el 19/7/2024.

GIL GARCÍA, Elizabeth: "Hacia el diseño de una inteligencia artificial garantista en el contexto tributario", en OLIVARES OLIVARES, Bernardo D. (Dir.): *La inteligencia artificial en la relación entre los obligados y la Administración Tributaria*. Madrid, AEDAF, 2022, pp. 75-94.

GÓMEZ REQUENA, José Ángel: "El empleo de la Inteligencia Artificial y el determinismo decisional", en OLIVARES OLIVARES, Bernardo D. (Dir.): *La inteligencia artificial en la relación entre los obligados y la Administración Tributaria*. Madrid, AEDAF, 2022, pp. 163-189.

HILDEBRANDT, Mirelle: "Code Driven Law. Scaling the Past and Freezing the Future", en MARKOU, Christopher y DEAKIN, Simon F. (eds.): *Critical Perspectives on Law and Artificial Intelligence.* Oxford, Hart Publishing, 2020, pp. 67-83. Disponible en https://bit.ly/3tEP1mY. Consultado el 10/10/2023.

IBFD OBSERVATORY ON THE PROTECTION OF TAXPAYERS' RIGHTS: *The 2023 IBFD Yearbook on Taxpayers' Rights*. Ámsterdam, IBFD, 2024, p. 48.

KAMINSKI, Margot E. y URBAN, Jennifer M.: "The Right to Contest AI". *Columbia Law Review* (2021), 121(7), pp. 1957–2048. Disponible en: https://www.jstor.org/stable/27083420. Consultado el 20/7/2024.

LOEVINGER, Lee: "Jurimetrics, the next step forward". *Jurimetrics Journal* (1971), 12(1), pp. 3-41. Disponible en: http://www.jstor.org/stable/29761220. Consultado el 19/7/2024.

MCCARTY, L. Thorne: "Reflections on 'Taxman': An Experiment in Artificial Intelligence and Legal Reasoning". *Harvard Law Review* (1977), 90(5), pp. 837-893. Disponible en: https://doi.org/10.2307/1340132. Consultado el 19/7/2024.

MOSQUEIRA-REY, Eduardo, HERNÁNDEZ-PEREIRA, Elena, ALONSO-RÍOS, David, BOVES BASCARÁN, José y FERNÁNDEZ LEAL, Ángel: "Human-in-the-loop machine learning: a state of the art", en *Artificial Intelligence Review* (2023), 56, pp. 3005–3054. Disponible en: https://doi.org/10.1007/s10462-022-10246-w. Consultado el 23/7/2024.

OCDE, *AI Principles*. OCDE (2024), París, sitio web. Disponible en: https://tinyurl.com/3zc7dtjr. Consultado el 23/7/2024.

OCDE: *Encuesta anual mundial*. París, OCDE, 2021. Disponible en: https://rb.gy/i0bn8p. Consultado el 5/7/2024.

OECD "Use of artificial intelligence, 2022: Percent of administrations", en *Tax Administration 2022: Comparative Information on OECD and other Advanced and Emerging Economies*. Paris, OECD Publishing, 2022. Disponible en: https://doi.org/10.1787/9b4d363d-en. Consultado el 5/7/2024.

OLIVER CUELLO, Rafael: "Big data e inteligencia artificial en la Administración tributaria". *Revista de Internet, Derecho y Política* (2021) nº 33, pp. 1-13. Disponible en: https://bit.ly/4f8fXiu. Consultado el 19/7/2024.

PÉREZ BERNABÉU, Begoña: "Decisiones automatizadas en la Administración Tributaria", en OLIVARES OLIVARES, Bernardo D. (Dir.): *La inteligencia artificial en la relación entre los obligados y la Administración tributaria. Retos ante la gestión tecnológica.* Madrid, AEDAF, 2022, pp. 145-162.

PONCE SOLÉ, Juli: "Inteligencia artificial, Derecho administrativo y reserva de humanidad: algoritmos y procedimiento administrativo debido tecnológico", en *Revista General de Derecho Administrativo* (2019), 50. Disponible en https://tinyurl.com/28bolu8u. Consultado el 23/7/2024.

RANGONE, Nicoletta: "Artificial Intelligence Challenging Core State Functions: A Focus On Law-Making And Rule-Making", en *Revista de Derecho Público: Teoría y método* (2023), 8, pp. 95-121. Disponible en: https://bit.ly/3Y7w9dP. Consultado el 20/7/2024.

RUIBAL PEREIRA, Luz: "Inteligencia Artificial, humanismo tecnológico e interpretación jurídica en el ámbito de un proceso contencioso-tributario", en SERRANO ANTÓN, Fernando (Dir.) *Inteligencia Artificial y Administración Tributaria: Eficiencia Administrativa y Defensa de los Derechos de los Contribuyentes.* Cizur Menor, Thomson Reuters Aranzadi, 2021, pp. 97-1

RUSSELL, Stuart J y NORWIG, Peter: *Artificial Intelligence: A Modern Approach.* 3ª edición, Upper Saddle River, Pearson Education, 2010. Disponible en: https://rb.gy/lp6pbu. Consultado el 19/7/2024.

SEN, Abhijeet: *Income Tax Department Sends Notices To Tax Evaders Using Artificial Intelligence*. India.com (2023), sitio web. Disponible en: https://bit.ly/4d8lljW. Consultado el 19/7/2024. Otro tanto ha ocurrido en Polonia y Estados Unidos.

"STATE V. LOOMIS", *Harvard Law Review* 130, nº 5 (2017), pp. 1530-1537. Disponible en: https://bit.ly/4d2BJCZ. Consultado el 19/7/2024.

STEDH del 14/5/2013, caso *N.K.M. v. Hungría* (App. nº 66529/11), párr. 54, 65. Disponible en https://tinyurl.com/2bawrdt5. Consultado el 23/7/2024.

SURDEN, Harry: "Artificial Intelligence and Law: an Overview". *Georgia State University Law Review* (2019), 35(4), pp. 1305-1337. Disponible en: https://bit.ly/46cstcy. Consultado el 19/7/2024.

SUSSKIND, Richard E.: "Expert Systems in Law: A Jurisprudential Approach to Artificial Intelligence and Legal Reasoning". *The Modern Law Review* (1986), *49*(2), pp. 168–194. Disponible en: http://www.jstor.org/stable/1096291. Consultado el 19/7/2024.

SUTTON, Sean: *How is AI & Machine Learning Revolutionizing the Tax Landscape*. Taina.com (2023), sitio web. Disponible en: https://bit.ly/3WsSdhF. Consultado el 19/7/2024.

TANZI, Vito y ZEE, Howell H.: "Tax Policy for Emerging Markets: Developing Countries". *National Tax Journal* (2000), 53(2), pp. 299–322. Disponible en: https://rb.gy/k4j0uw. Consultado el 19/4/2024.

TAYLOR, Isaac: "Justice by Algorithm: The Limits of AI in Criminal Sentencing". *Criminal Justice Ethics* (2023), 42(3), pp. 193–213. Disponible en: https://doi.org/10.1080/0731129X.2023.2275967. Consultado el 19/7/2024.

TORRECILLA SALINAS, Carlos *et al*.: "¿Para qué sirve la Inteligencia Artificial en el sector público? Casos de uso y perspectivas de aplicación", en GAMERO CASADO, Eduardo (Dir.) y PÉREZ GUERRERO, Francisco (coord.), *Inteligencia Artificial y sector público: retos, límites y medios*. Valencia, Tirant Lo Blanch, 2023, pp. 73-91.

VOERMANS, Wim: *Computer-assisted legislative drafting in the Netherlands: the LEDA system*. Tillburg, S/E, pp. 1-10. Disponible en: https://tinyurl.com/27q5aclj. Consultado el 23/7/2024

WEFFE, Carlos E.: "Mandatory Disclosure Rules and Taxpayers' Rights: Where Do We Stand?". *International Tax Studies* (2021), 4(1), pp. 1-17. Disponible en: https://bit.ly/3z1d-H6R. Consultado el 19/4/2024.

WEFFE, Carlos E.: "The Right to Be Informed: The Parallel between Criminal Law and Tax Law, with Special Emphasis on Cross-Border Situations", *World Tax Journal* (2017), 9(3), pp. 430-472. Disponible en: http://bit.ly/2yv0Gnz. Consultado el 19/7/2024.

WEFFE, Carlos E.; "Notas Introductorias al Análisis Económico del Derecho Penal Tributario", en SALAVERRÍA, José Getulio (Dir.), *XIII Jornada Aníbal Dominici. Análisis Económico del Derecho en Homenaje al Dr. Humberto Romero-Muci.* Caracas, Universidad Católica Andrés Bello, 2023 (a), pp. 335-366. Disponible en: https://bit.ly/47IBp9k. Consultado el 19/7/2024.

WEFFE H., Carlos E.: "Inteligencia Artificial, proceso de creación normativa y *Tax Administration* 3.0. Una primera aproximación al riesgo del 'legalismo computacional' tributario", en OLIVER CUELLO, Rafael (Dir.): *El Derecho, la Empresa y la Comunicación en la sociedad de la información*. J.M. Bosch Editor, Barcelona (España), 2023 (b), pp. 525-550.

Capítulo 17

¿SUEÑAN LOS INVENTORES CON EXAMINADORES ELÉCTRICOS? LA IA EN UN CONTEXTO ADMINISTRATIVO APLICADO: EL EXAMEN SOBRE NOVEDAD Y ACTIVIDAD INVENTIVA DEL PROCEDIMIENTO DE CONCESIÓN DE PATENTES.

José Antonio Gil Celedonio.
Ministerio de Industria y Turismo.

1. INTRODUCCIÓN

La influencia de la Inteligencia Artificial (IA, en adelante) en el debate público y en los debates académicos parece haber venido para quedarse: en la actualidad, no hay disciplina que no analice (o directamente incorpore) este relevante conjunto de tecnologías en sus discusiones internas. El ámbito jurídico no ha permanecido ajeno a este proceso: a día de hoy, operadores jurídicos de todo tipo y nivel (en el sector público y en el ámbito privado) ya incorporan la IA en sus actividades diarias y planes de acción, y han internalizado sus conceptos en sus estrategias. Su aparición en los debates políticos y jurídicos se ha incrementado exponencialmente a partir de la publicación de la Estrategia para el Mercado Único Digital de 2015 y, especialmente, del documento de la Comisión Europea (en adelante, la Comisión) de 2018 *Artificial Intelligence for Europe*.[1] Esta tendencia no ha hecho sino acelerarse desde abril de 2021[2], cuando la Comisión hizo público el borrador de acto legislativo conocido como "*AI Act*", la primera propuesta legislativa de carácter horizontal en la UE en la materia[3], que exponía como pretensión principal la de establecer reglas sobre este conjunto de tecnologías en el ya denso *acquis communautaire*.

La IA y sus tecnologías asociadas, estadios actuales de un amplio proceso de digitalización de la sociedad cuyo inicio puede observarse desde las últimas décadas del siglo XX[4], no son nuevas en sí mismas[5], pero son aceleradores del cambio en términos schumpeterianos, y, con su capacidad de "*destrucción creativa*", generan innovaciones que afectan profundamente a algunas instituciones sociales, económicas y jurídicas ya consolidadas y empujan a su

1 COM (2018) 237 final.

2 No en vano, un estudio ha mostrado que la frecuencia de la aparición del término "IA" era casi nula en 2014, pero apareció en un 3% del total de los documentos publicados por la UE en 2021, lo que da muestra de su naturaleza estratégica. Puede verse en KRAKUP, Troels, HORST, Maja: "European artificial intelligence policy as digital single market making", *Big Data and Society* (2023), January-June, 2023, 1-14, p. 5.

3 En el momento en que se escribe este texto, esta propuesta, ampliamente modificada tras las negociaciones interinstitucionales entre el Consejo y el Parlamento Europeo (concluidas bajo la Presidencia Española del Consejo de la Unión Europea a finales de 2023), ha visto su versión final adoptada y publicada en el Diario Oficial de la Unión Europea del día 12 de julio de 2024 como Reglamento (EU) 2024/1689, del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) nº300/2008, (UE) nº 167/2013, (UE) nº 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). Dado el objeto de este trabajo, así como su limitada extensión, no se podrá abordar en específico el fuerte impacto que es probable que esta normativa tenga sobre múltiples campos públicos y privados y, en particular, sobre nuestro objeto de estudio.

4 Para un análisis en detalle, vid. GIL CELEDONIO, José Antonio: "Digitalización, inteligencia artificial y sus impactos jurídicos: una aproximación al caso de la propiedad industrial" en CANDELARIO MACÍAS, María Isabel (dir.): *La propiedad industrial en tiempos del COVID-19*. Tirant Lo Blanch, 2022, pp. 17-25.

5 CERRILLO i MARTÍNEZ, Agustín: "El derecho para una inteligencia artificial centrada en el ser humano y al servicio de las instituciones", *Revista de Internet, Derecho y Política* (2020), nº 30, pp. 2-3; CAPDEFERRO VILLAGRASA, Óscar: "La inteligencia artificial del sector público: desarrollo y regulación de la actuación administrativa inteligente en la cuarta revolución industrial", *Revista de Internet, Derecho y Política* (2020), nº 30, pp. 3-5.

modificación y/o destrucción, e incluso a la necesidad de buscar nuevos equilibrios ante las disrupciones que crean, con ganadores y perdedores[6]. Este fenómeno impacta de lleno en el ordenamiento jurídico, generando nuevas facultades que han de incorporarse al objeto de los derechos fundamentales clásicos, dando nuevas perspectivas para la eficacia de los mismos, tanto frente a los poderes públicos como en las relaciones entre individuos, e incluso generando el nacimiento de algunos nuevos[7], poniendo en riesgo elementos clave de nuestros sistemas constitucionales y administrativos y cuestionando la capacidad de la Constitución para ordenar jurídica y efectivamente la sociedad digital[8].

De entre las diferentes ramas del ordenamiento jurídico, el derecho de la propiedad intelectual e industrial, en términos generales, y el derecho de patentes, en particular, no son extraños a la ya mencionada tendencia, debido a las más que obvias relaciones entre dichos sectores del ordenamiento jurídico y las últimas fronteras de la innovación tecnológica. Hasta el momento, la interacción entre el derecho de patentes y la IA ha sido abordada desde dos diferentes puntos de vista[9]: el primer enfoque aborda el debate en relación el objeto de la protección de la patente, la materia patentable, analizando pues la posibilidad de que se concedan patentes sobre invenciones relacionadas con la IA o que la contengan (lo que se conoce como invenciones implementadas por ordenador[10]). El segundo, mucho más complejo y también controvertido, trata de poner luz sobre la titularidad de la patente en el caso de invenciones generadas por máquinas o sistemas de IA y, como corolario, el reconocimiento (o no) de estos sistemas de IA como inventores (algo que ha dejado de ser teórico tras la irrupción del caso DABUS)[11], con una controvertida y potencial atribución de subjetividad, de una manera o de otra[12].

6 KERBER, Wolfgang: "Digital revolution, institutional coevolution and legal innovations", *European Bussiness Law Review* 34 (2023), nº 6, pp. 995-996.

7 PRESNO LINERA, Miguel Ángel: "Teoría General de los Derechos Fundamentales e Inteligencia Artificial: una aproximación", *Revista Jurídica de Asturias* (2022) nº 45, pp. 59-65.

8 SÁNCHEZ BARRILAO, Juan Francisco: "Inteligencia Artificial y fuentes del derecho", *Revista española de Derecho Constitucional Europeo* (2023), núm. 39, pp. 135 y ss.

9 EUROPEAN COMMISSION, DIRECTORATE-GENERAL FOR COMMUNICATIONS NETWORKS, CONTENT AND TECHNOLOGY, HARTMANN, Christian, ALLAN, Jaqueline, et al.: *Trends and developments in artificial intelligence – Challenges to the intellectual property rights framework – Final report*, Publications Office of the European Union, 2020, p. 100-113. https://data.europa.eu/doi/10.2759/683128

10 MCLAUGHIN, Michael: "Computer-generated inventions", *Social Science Research Network Electronic Journal* (2018) nº 15, pp. 1-32; GALLEGO SÁNCHEZ, Esperanza: "La patentabilidad de la inteligencia artificial. La compatibilidad con otros mecanismos de protección", *La Ley Mercantil (2019)*, nº 59.

11 GIL CELEDONIO, José Antonio: *opus cit. 2022,* pp. 39-47.

12 ABBOTT, Ryan: "I think, therefore I invent: creative computers and the future of patent law", *Boston College Law Review* (2016) Vol. 57, pp. 1079-1126; MAROÑO GARGALLO, María del Mar: "El concepto de inventor en el derecho de patentes y los sistemas de Inteligencia Artificial", *Cuadernos de Derecho Transnacional* (2020), vol. 12, nº 2, pp. 510-526; SÁNCHEZ GARCÍA, Luz: "Propuesta de una nueva categoría de inventor para el sistema de patentes español. El inventor artificial", *Comunicaciones en Propiedad Industrial y derecho de la competencia* (2019), nº 86, pp. 77-90.

En esta comunicación, pretendemos traer a colación un tema no demasiado estudiado hasta el momento, la intersección entre el derecho administrativo y el derecho de patentes: hasta qué punto las herramientas de IA pueden ser utilizadas en el contexto del procedimiento de concesión de patentes. Asumiendo la premisa de que las decisiones administrativas no serán adoptadas ya exclusivamente por humanos, sino por humanos con ayuda de máquinas o, directamente, sin intervención humana[13], analizaremos cómo pueden introducirse este tipo de herramientas en este ámbito de la actuación administrativa de las Oficinas Nacionales de Propiedad Industrial (o sus homólogas regionales) sin que esto impacte de forma significativa en los pilares más básicos de nuestros sistemas de patentes, así como las consecuencias que de ello podrían derivarse.

2. LAS OFICINAS DE PROPIEDAD INDUSTRIAL Y LA INTELIGENCIA ARTIFICIAL EN SUS PROCEDIMIENTOS Y TAREAS ADMINISTRATIVAS.

Las Oficinas de Propiedad Industrial (cuyo referente en España es la Oficina Española de Patentes y Marcas, O.A.[14]) son entidades administrativas que desarrollan una actividad de servicio público dirigida tanto a los ciudadanos particulares como a las empresas, y, por ello, el cauce por el que se materializa dicha actividad (el procedimiento de concesión de patentes, *inter alia*) es el procedimiento administrativo, pero con una naturaleza especial frente al procedimiento administrativo común. Como todo procedimiento administrativo, comprende varios pasos consecutivos, con un mayor o menor grado de complejidad, que permiten a una Oficina de Propiedad Industrial llegar a una decisión final informada en algo tan relevante como la atribución de un derecho de propiedad (intangible), materializando la decisión en el acto administrativo correspondiente, concediendo o denegando el derecho, ya que el registro del derecho goza de carácter constitutivo.

En tanto parte del sector público, estos organismos llevan ya tiempo analizando cómo utilizar las diferentes herramientas de IA para acelerar los procedimientos de concesión, así como para mejorar la solidez de los mismos. Lideradas desde 2018 por la Organización Mundial para la Propiedad Intelectual (OMPI, en adelante), oficinas de todo el mundo han manifestado su voluntad de introducir, o, cuando menos, explorar la introducción de herramientas asistidas por IA en sus rutinas laborales diarias. Tal y como recoge un documento de OMPI, "*las herramientas de IA han demostrado ser eficaces y útiles en aquellas áreas que requieren rutinas y trabajos de ejecución estricta de normas específicas, tales como los exámenes de formalidades, la determinación y el reparto de los símbolos de clasificación más relevantes, así como para la distribución interna de expedientes de solicitud a las unidades de examen competentes*", señalando que las áreas en las que el

13 PONCE SOLÉ. Juli: "Inteligencia Artificial, decisiones administrativas discrecionales totalmente automatizadas y alcance del control judicial: ¿indiferencia, insuficiencia o deferencia? *Revista de Derecho Público: Teoría y Método* (2024) Vol. 9, p 171 y ss.

14 Vid., al respecto, el Real Decreto 1270/1997, de 24 de julio, por el que se regula la Oficina Española de Patentes y Marcas.

uso de herramientas asistidas por IA están especialmente extendidas son, esencialmente, las siguientes: herramientas lingüísticas para la traducción de documentos, la distribución interna de expedientes, verificación de capturas de datos de documentos en papel (una vez digitalizados); clasificación automática o pre-clasificación de elementos contenidos en solicitudes y en las búsquedas de patentes o de marcas.[15] Este camino parece prometedor en un sector del que se ha dicho que, tradicionalmente, ha estado caracterizado por grandes cantidades de documentos en papel, laboriosas búsquedas manuales y una administración muchas veces ineficiente y tendente al error.[16]

En el ámbito europeo puede observarse también esta corriente: en 2020, la Comisión ya llamó la atención sobre las posibilidades de aplicación de herramientas de IA en el ámbito de la propiedad intelectual e industrial. En su Comunicación aprobando su Plan de Acción en materia de Propiedad Intelectual, redactado en línea con el contexto general de la transición digital, la Comisión subrayó la relevancia de la revolución tecnológica que conllevan la IA y su amplio conjunto de tecnologías para coadyuvar a una mejor protección de los activos intangibles, mejorando la efectividad de los sistemas ya existentes. En sus propias palabras, "*es necesario tener en cuenta el potencial ofrecido por nuevas tecnologías como la IA y la cadena de bloques para incrementar la efectividad de nuestros sistemas de propiedad intelectual e industrial*"[17]. En particular, "*las nuevas tecnologías pueden ayudar a facilitar la protección de la propiedad intelectual e industrial, por ejemplo, mediante la aceleración de las búsquedas de novedad y de los procedimientos de registro, la mejora de la transparencia, una mayor distribución de los regalías por licencias y un mejor abordaje de la lucha contra la falsificación y la piratería*"[18]. De una manera más precisa, una reciente comunicación de la Comisión sobre medidas para combatir la falsificación considera que "*los sistemas de IA, tales como sistemas automatizados de reconocimiento de contenido o algoritmos de aprendizaje automático, pueden ser utilizados para reconocer productos falsificados o pirateados o tendencias relacionadas con la promoción o la distribución de dichos productos, y tienen el potencial de convertirse en tecnologías clave en la lucha contra las actividades infractoras de los derechos de propiedad intelectual*"[19].

Sin perjuicio de las múltiples y obvias ventajas de la aplicación de estas herramientas en determinadas áreas del trabajo administrativo de las Oficinas de Propiedad Industrial, ha de formularse una serie de preguntas primordiales a fin de evitar caer en el conocido como sesgo de automatización[20]: ¿es todo lo técnicamente posible recomendable o viable en términos administrativos? ¿sirve el instrumento para los fines pretendidos? ¿superan las ventajas a los riesgos en relación con el procedimiento más complejo que todas estas Oficinas abordan, como es el examen de patentabi-

15 Doc. WIPO/IP/GE/19/1, pp. 2-3.

16 VANHERPE, Josephine: "AI and IP- Great Expectations", en DE BRUYNE, Jan, VANLEENHOVE, Cedric (eds.): *Artificial Intelligence and the Law, 2nd ed.* KU Leuven Centre for IT&IP Law Series, 2023, p. 236.

17 COM (2020) 760 final, pp. 2-3

18 COM (2020) 760 final, p. 7.

19 COM (2024) 1738 final, p. 10.

20 CERRILLO I MARTÍNEZ, Agustí: "¿Son fiables las decisiones de las Administraciones públicas adoptadas por algoritmos? *European Review of Digital Administration and Law- Erdal* (2020), Vol. 1, issue 1-2, p. 19.

lidad? En definitiva ¿Cuán lejos y en qué medida pueden utilizarse estas herramientas de IA para la determinación de los requisitos más relevantes para la concesión de patentes?

3. EL EXAMEN DE NOVEDAD Y ACTIVIDAD INVENTIVA COMO TRÁMITES CUALIFICADOS DEL PROCEDIMIENTO DE CONCESIÓN.

Para que una patente sea concedida la invención debe satisfacer, entre otros, tres requisitos, conocidos como requisitos secundarios de patentabilidad: la invención debe ser nueva, debe presentar actividad inventiva y ser susceptible de ser aplicación industrial.[21] Estos principios básicos tienen gran reconocimiento en el derecho de patentes, y se han visto internacionalmente confirmados con la adopción del Acuerdo de la Organización Mundial del Comercio sobre los Aspectos de los Derechos de Propiedad Intelectual relacionados con el Comercio (en adelante, ADPIC) de 1994 y su subsiguiente entrada en vigor para todos los Estados parte de dicha Organización Internacional. De este modo, el artículo 27.1 de ADPIC establece claramente que "*las patentes podrán obtenerse por todas las invenciones, sean de productos o de procedimientos, en todos los campos de la tecnología, siempre que sean nuevas, entrañen una actividad inventiva y sean susceptibles de aplicación industrial*". Es por ello que las Oficinas de Propiedad Industrial están obligadas a examinar el cumplimiento de dichos requisitos sustantivos en cada invención descrita en una solicitud de patente sometida a su consideración mediante la conducción de un procedimiento administrativo concreto. En este sentido, todas las formalidades para examinar la novedad, la actividad inventiva y la aplicación industrial son trámites clave en el procedimiento de concesión del derecho.[22] Los actos administrativos resultantes de la evacuación de estos trámites, como es el caso del informe sobre el estado de la técnica y su opinión escrita[23] o el

21 PILA, Justine, TORREMANS, Paul: *European Intellectual Property Law,* 2nd ed., Oxford University Press, 2019, p. 162; BENTLY, Lionel, SHERMAN, B., GANGJEE, Dev, JOHNSON, Phillip, *Intellectual Property Law*, 5th ed., Oxford University Press, 2018, pp. 458 y ss.

22 La concesión eficaz y jurídicamente consistente de los títulos de propiedad industrial, de manera expeditiva y eficiente, es quizá la principal tarea de toda Oficina de Propiedad Industrial no solo en el ámbito de la protección de las innovaciones de carácter técnico mediante los derechos de patente y modelo de utilidad (entre otros) sino también en el ámbito de los derechos sobre signos distintivos (marcas, normales comerciales y, eventualmente, signos de calidad como las denominaciones de origen y las indicaciones geográficas) y aquellos que protegen innovaciones formales o de carácter estético (diseño industrial). Estas tareas son esenciales en tanto, además, el nacimiento del derecho se somete a su inscripción registral, que reviste carácter constitutivo. No obstante, dentro del catálogo de tareas que tienen encomendadas, destacan también servicios más innovadores que se vienen ofreciendo desde tiempos más recientes a los usuarios y a la ciudadanía, en general (como los servicios de información tecnológica o las actividades que fomentan la observancia y la promoción de este conjunto de derechos). En este sentido, puede verse GIL CELEDONIO, José Antonio: "La Inteligencia artificial aplicada en el sector público: el caso de las Oficinas de Propiedad Industrial y, en particular, el de la Oficina Española de Patentes y Marcas, O.A.", *Comunicaciones en Propiedad Industrial y Derecho de la Competencia* (2023), nº 99, pp. 77-81; CASADO CERVIÑO, Alberto: "Globalización y propiedad industrial: el escenario emergente", *Comunicaciones en Propiedad Industrial y Derecho de la Competencia* (2008), nº 49, pp. 69-72.

23 Art. 36 de la Ley 24/2015, de 24 de julio, de Patentes y arts. 26 a 32 del Real Decreto 316/2017, de 31 de marzo, por el que se aprueba el Reglamento para la ejecución de la Ley 24/2015, de 24 de julio, de Patentes.

resultado del examen sustantivo[24] son, sin duda, trámites cualificados (a través de los que la autoridad competente decide directa o indirectamente el fondo del asunto) mediante los que se ejercitan potestades discrecionales, en el sentido del artículo 35.1.i) de la Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas.

En el derecho europeo de patentes (al que se adhiere la legislación española en su totalidad), se considera que una invención es nueva cuando no forma parte del estado de la técnica, concepto que, a su vez, parece definido como "*todo aquello hecho público por medio de una descripción oral o escrita, por el uso o por cualquier otra vía, antes de la fecha de presentación de una solicitud de patente europea*"[25]. Por tanto, el examen para discernir si una invención debería ser considerada nueva o no es determinado *per comparationem*, realizando un ejercicio de comparación entre lo que se describe en las reivindicaciones plasmadas en la solicitud y lo comprendido en el estado de la técnica, que alcanza esencialmente todo, desde principios científicos a reglas técnicas, métodos, conocimientos generales y específicos, etc…[26]. Estas operaciones requieren de tiempo y experiencia, en tanto el estado de la técnica es una vasta cantidad de datos muy diversos. Tal y como establecen las Directrices de Examen de la Oficina Europea de Patentes (EPO, en adelante) en términos muy precisos, para la determinación de la novedad de la invención "*ha de decidirse qué contenido material susceptible de protección ha sido puesto en conocimiento al público mediante cualquier divulgación anterior y, por tanto, forma parte del estado de la técnica. En este contexto, no son solo ejemplos parciales, sino el contenido completo del documento que se considera como anterioridad es el que ha de tenerse en cuenta*"[27].

Por consiguiente, ¿no sería útil tener herramientas de IA que, usando su amplia capacidad para la explotación de grandes conjuntos de datos, pudiese analizar y comparar un conjunto de reivindicaciones dado contra una base de datos específica, tan grande como sea posible? Nuestra valoración es afirmativa. El uso de este tipo de tecnologías podría ser beneficioso para todas las partes involucradas en este estadio del procedimiento: dado que, como se mencionó anteriormente, las capacidades de procesamiento de información de la herramienta de IA y un algoritmo de comparación específico harían que las búsquedas de anterioridades fuesen más extensas y refinadas, y podrían aportar una base muy sólida para el trabajo posterior del examinador de patentes (o el panel de examen, compuesto por va-

24 Art. 39 de la Ley 24/2015, de 24 de julio, de Patentes y art. 34 del Real Decreto 316/2017, de 31 de marzo, por el que se aprueba el Reglamento para la ejecución de la Ley 24/2015, de 24 de julio, de Patentes.

25 Art. 54.2 del Convenio sobre concesión de Patentes Europeas, hecho en Múnich el 5 de octubre de 1973. A los efectos del texto, hablaremos del CPE (Convenio de Patentes Europeas). En relación con la fecha de presentación, y para evitar digresiones innecesarias, debe entenderse también la fecha de prioridad puesta de manifiesto por el solicitante, haciendo uso del derecho que le atribuye el artículo 4 del Convenio de París para la Protección de la Propiedad Industrial de 1883, en su versión enmendada el 28 de septiembre de 1979.

26 FERNÁNDEZ NÓVOA; Carlos, OTERO LASTRES, José Manuel; BOTANA AGRA, Manuel: *Manual de la propiedad industrial*, 3ª ed., 2017. Marcial Pons; Madrid, pp. 118-120.

27 EUROPEAN PATENT OFFICE, *Guidelines for examination*. Ed. 2024, Part G, Chapter VI, pp. 801-805. https://link.epo.org/web/legal/guidelines-epc/en-epc-guidelines-2024-hyperlinked.pdf.

rios examinadores), lo que no solamente mejoraría la calidad de dicha búsqueda, sino que también liberaría tiempo de trabajo que podría ser dedicado a otras actividades de mayor valor añadido, dentro o fuera del procedimiento de concesión. Desde el punto de vista de los usuarios y/o del solicitante, se podrían reducir drásticamente los plazos asociados del procedimiento, dando por tanto más tiempo al solicitante a decidir, por ejemplo, si proseguir con el procedimiento o si extenderlo a otras jurisdicciones para obtener protección. En definitiva, el resultado sería mejor que el de un humano sin que los potenciales perjuicios (prácticamente inexistentes, en nuestra opinión) superen a los riesgos aceptables de uso, en clara puesta en práctica del principio de cautela o precaución[28].

No obstante, la situación no es tan sencilla si hablamos del examen de la actividad inventiva, en tanto se pasa de analizar diferencias cuantitativas entre la invención y el estado de la técnica a una apreciación verdaderamente cualitativa[29]. Si el examen de novedad es, fundamentalmente, un ejercicio comparativo de carácter retrospectivo, el juicio sobre la actividad inventiva es una hipótesis a seleccionar entre varias tras la realización de un ejercicio de prospectiva. Tal y como se reconoce en el artículo 56 del CPE, una invención conlleva actividad inventiva si, teniendo en cuenta el estado de la técnica, no resulta obvia a una persona experta en la materia. Este requisito persigue un claro objetivo: las patentes deben proteger solamente aquellas invenciones que presentan un paso adelante suficiente desde el estado de la técnica al que la invención pertenece, a fin de evitar conceder derechos de exclusiva sobre meros avances tecnológicos que serían obvios para un especialista y quizá también para la población general.

En términos prácticos, analizar la existencia de actividad inventiva supone otra compleja operación que puede ser resumida en los pasos siguientes: en primer lugar, deben identificarse los caracteres técnicos esenciales de la invención cuya patente se busca obtener; en segundo lugar, debe determinarse tanto la fecha de presentación como el estado de la técnica y, en tercer lugar, se examina si, a la vista del estado de la técnica en dicho momento, la invención sería obvia o no para el experto en la materia. La pieza clave, por tanto, resulta ser la apreciación por parte del denominado como experto en la materia (*skilled addressee*). Esta *persona* es la encargada de llevar a cabo los trámites administrativos necesarios para determinar si la invención conlleva actividad inventiva, esto es un, examinador de patentes profesionalmente dedicado a ello en el ámbito de una Oficina de Propiedad Industrial[30]. Las Directrices de

28 CIERCO SIERRA, C: "El principio de precaución: reflexiones sobre su contenido y alcance en los Derechos comunitario y español", *Revista de Administración Pública* (2004), núm. 163, pp. 73-126.

29 BENTLY, Lionel, SHERMAN, Brad, GANGJEE, Dev, JOHNSON, Phillip: *opus cit.*, p. 578 y ss.

30 En el caso español, los examinadores de patentes son funcionarios de carrera de la Escala de Titulados Superiores de Organismos Autónomos del Ministerio de Industria y Turismo, especialidad de Propiedad Industrial, colectivo funcionarial adscrito al Subgrupo A1, según la clasificación establecida por la disposición transitoria tercera del Real Decreto Legislativo 5/2015, de 30 de octubre, por el que se aprueba el texto refundido de la Ley del Estatuto Básico del Empleado Público. Se considera que este colectivo funcionarial participa directa o indirecta en el ejercicio del poder público o en funciones que tienen por objeto la salvaguarda de los intereses del Estado o de las Administraciones públicas, en tanto el anexo del Real Decreto 543/2001, de 18 de mayo, sobre acceso al empleo público de la Administración General del Estado y sus Organismos públicos de nacionales de

Examen de la EPO lo definen como "*en un campo específico de la tecnología, aquel que posee un conocimiento medio y habilidad… y está al tanto de lo que es un conocimiento general común de la técnica en el momento relevante, así como se presume que ha tenido acceso a todo lo comprendido en dicho estado de la técnica*".[31]

Por tanto, en este contexto, el examinador de patentes que trabaje en una Oficina de Patentes de acuerdo con una relación profesional estable, dedicado y conocedor de su campo técnico, es el encargado de hacer uso de las diferentes herramientas puestas a su disposición (incluyendo herramientas tecnológicas) para examinar en profundidad si la invención presenta dicha actividad inventiva. Pero, sobre todo, es la persona a la que se confía la conducción de complejas operaciones intelectuales de comparación entre lo que se ha divulgado por el solicitante en las reivindicaciones y lo que podría constituir la frontera tecnológica inmediatamente posterior al estado de la técnica actual, algo que debe concebir/imaginar desde su conocimiento técnico. En consecuencia, este tipo de valoraciones, incluso produciéndose con seguimiento estricto de directrices y métodos, puede llevar a resultados incoherentes e incluso impredecibles sobre la presencia o no de la actividad inventiva ante un mismo caso, según quién y cómo se examine.[32]

Pero imaginemos, por un momento, que a dicho examinador de patentes se le da una nueva herramienta tecnológica que, como una herramienta de IA, pudiese llevar a cabo parte o la totalidad de las tareas respecto de este examen. Basándonos en su capacidad para analizar grandes cantidades de datos de una manera muy sofisticada, la herramienta poseería una poderosa habilidad para predecir o visionar la siguiente frontera incremental de un campo técnico específico, lo que se ha dicho elevaría de manera clara la barrera para que la invención pudiese cumplir con el requisito de la actividad inventiva, reduciendo en mucho la franja de invenciones que podrían, en consecuencia, ser patentables.[33] Se ha apuntado, en la misma línea, que si el estándar de juicio humano fuese reemplazado por el denominado "*estándar inventivo de la máquina*", el punto de referencia actual para la patentabilidad se elevaría dramáticamente, en tanto estos "examinadores eléctricos" serían significativamente más inteligentes que los ya mencionados expertos en la materia de la actualidad y, lo que es aún más problemático, serían capaces de considerar mucho más estado de la técnica, por la que la inferencia final, potencialmente, elevaría el estándar analítico, imponiendo una muy alta barrera de entrada a cualquier invención.

En un sistema caracterizado por una innovación descentralizada en muchos puntos a lo largo y ancho del mundo, en el que muchas entidades públicas y privadas trabajan al mismo tiempo, este cambio drástico en la barrera de la actividad inventiva y, en consecuencia, de

otros Estados a los que es de aplicación el derecho a la libre circulación de trabajadores reserva a este colectivo para candidatos con nacionalidad española.

31 EUROPEAN PATENT OFFICE, *ibid*. Part G, Chapter VII, p 817.

32 ABBOTT, Ryan: "Everything is obvious", *66 UCLA Law Review (2019)*, 2, p. 7.

33 ÍÑIGUEZ, Pilar: "Intellectual Property Rights, Artificial Intelligence and Big Data: future perspectives", *Actas de Derecho Industrial (*2021*), nº 41*, p. 119.

la patentabilidad en sí, impactaría severamente en el sistema de patentes en su conjunto. Los actuales polos científicos y tecnológicos generan multitud de innovaciones prácticas de naturaleza derivada[34], que, sin ser consideradas revolucionarias, pero con suficiente actividad inventiva, actualmente generan patentes pero no podrían ser patentadas, de seguirse este criterio. Esto minaría un sistema de patentes que, en la actualidad, parece satisfacer de manera adecuada las necesidades de la sociedad gracias a su papel de fomento de la innovación. Diferentes autores han hablado de la necesaria reserva de humanidad, esto es, la decisión humana de excluir el uso de la IA de ciertos ámbitos por considerarlo inadecuado, como el caso de las decisiones administrativas resultado del ejercicio de potestades discrecionales[35]. El anterior análisis parece apuntar que, en este caso, nos encontraríamos con uno de esos ejemplos de no idoneidad de una herramienta de IA para una actuación concreta.

4. CONCLUSIONES (Y UN CAVEAT FINAL)

Tras la publicación del Plan de Acción en materia de propiedad intelectual e industrial de la Comisión que se mencionó anteriormente, el Parlamento Europeo respondió con la aprobación de una Resolución en la cual puede leerse lo se siguiente: "…*se subraya que la IA y sus tecnologías conexas usadas en el procedimiento de concesión o registro de los derechos de propiedad intelectual e industrial no puede ser un sustituto de la revisión humana llevada a cabo caso por caso a fin de asegurar la calidad y la justicia de las decisiones*".[36] En efecto, el uso de este tipo de herramientas debe poner la visión humana en el centro, de tal modo que otros objetivos públicos como la calidad o la justicia (además de otros) no queden sacrificados en el altar de la eficacia. El derecho de patentes, obviamente, no es una excepción al respecto. No podemos permitirnos no utilizar las muchas ventajas de estas tecnologías, sino que, al contrario, deben estar disponibles para su uso con unos límites claramente orientados: debe encontrarse un equilibrio proporcional entre objetivos igual de meritorios, aunque difíciles de reconciliar, incluyendo siempre los valores de justicia, rendición de cuentas y transparencia[37]. Como se ha apuntado de manera aguda, el debate no puede ser regular o no regular para evitar riesgos, en tanto no regular es un riesgo en sí mismo.[38]

34 VAN CAENEGEM, William: *Intellectual Property Law and innovation*. Cambridge University Press, 2007, pp. 17-18.

35 PONCE SOLÉ, Juli: "Límites jurídicos de la toma de decisiones discrecionales automatizadas mediante Inteligencia Artificial: Racionalidad, Sabiduría y necesaria reserva jurídica de humanidad en el ámbito digital", *Revista General de Derecho Administrativo* (2024), núm. 66.; CERRILLO i MARTÍNEZ, Agustí: *opus cit.* (2020), p. 25.

36 2021/2007 (INI), *European Parliament Resolution of 11 November 2021 on an intellectual property action plan to support the EU's recovery and resilience*. https://www.europarl.europa.eu/doceo/document/TA-9-2021-0453_EN.html.

37 CALO, Ryan: "Artificial Inteligence Policy: a primer and roadmap", *University of Bologna Law Review* (2018) *3*(2), p. 191.

38 KOULU, Riikka, HIRVONEN, Hanne, SANKARI, Suvi, HEIKKINEN, Tatjaan: "Artificial Intelligence and the Law: Can and Should We Regulate AI Systems?", en: *Edward Elgar Handbook on Law and Technology*, Edward Elgar, 2023, p. 423.

Habida cuenta de que los derechos de propiedad intelectual e industrial están protegidos bajo el paraguas del artículo 17 de la Carta de Derechos Fundamentales de la UE, así como de la interacción entre estos derechos y otros muy relevantes derechos fundamentales más conocidos, resulta claro que la tecnología debe cumplir un propósito específico: las herramientas de IA son instrumentos orientados a la satisfacción de ciertas necesidades y para ayudar a alcanzar las aspiraciones de nuestras sociedades. Por ello, no deberíamos aceptar acríticamente cambios tecnológicos que no podemos manejar en el medio o largo plazo. El enfoque humanista es necesario para tratar este conjunto de tecnologías también en el ámbito del derecho de la propiedad intelectual e industrial, y, especialmente, en sus relevantes dimensiones prácticas: la decisión humana final no puede ser reemplazada salvo que, bajo una evaluación del riesgo adecuado, se demuestre que este es inexistente.

La pretensión del sistema de patentes no es otra que reconciliar la atribución de los derechos de propiedad del inventor con la necesidad pública de impulsar la innovación tecnológica y su distribución, un equilibrio delicado que es fácil de erosionar, pero difícil de mantener. Sin un sistema de patentes equilibrado no habría incentivos a la innovación, un escenario en el cual las futuras posibilidades de la humanidad se verían dañadas para siempre, ya que no existiría la tutela jurídica que es presupuesto ineludible para que se puedan dedicar a la investigación y a la innovación los adecuados recursos humanos y financieros[39]. No obstante, el elemento positivo de todo análisis prospectivo sobre potenciales escenarios es que éstos pueden ser evitados, siempre y cuando se tomen las acciones adecuadas. Por ello, la propuesta aquí no es abandonar el uso de las herramientas y sistemas de IA, sino utilizarlas de manera inteligente y con límites. Se han analizado algunos de ellos en el contexto del complejo procedimiento para la concesión de patentes, pieza clave de los modernos sistemas de protección de la propiedad industrial. Es esa la línea pergeñada por el ya mencionado Reglamento sobre Inteligencia Artificial, o la también reciente Convención Marco sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho, del Consejo de Europa.

En su libro de 1968, titulado muy sugerentemente "*¿Sueñan los androides con ovejas eléctricas?*"[40], el escritor Phillip K. Dick expresó lo siguiente: "*Se había preguntado, como casi todos en un momento u otro, por qué precisamente los androides se agitaban impotentes al afrontar el test de medida de la empatía. Era obvio que la empatía sólo se encontraba en la comunidad humana, aunque se podía hallar cierto grado de inteligencia en todas las especies, hasta en los arácnidos.*" En nuestra opinión, es una clara alegoría que ilustra los límites entre las capacidades humanas y las capacidades de cualquier tipo de instrumento al servicio de la humanidad. Como se ha dicho acertadamente, la IA y los humanos adoptan decisiones de forma distinta en el momento actual y esto debería conducir a una limitación del uso de estas herramientas de IA, de acuerdo con ya mencionado principio de cautela o precaución[41], sin olvidar que, en todo caso, la inteligencia más valiosa

39 MASSAGUER FUENTES, José: "La propiedad industrial: balance y perspectivas", *Actas de Derecho Industrial y derecho de la competencia* (1998), p. 97.

40 Posteriormente fue rebautizado con el añadido "Blade Runner", en clara muestra estrategia comercial basada en el *rebranding,* tras el estreno de la película homónima de Ridley Scott (estrenada en 1982).

41 PONCE SOLÉ, Juli: *opus cit.* 2024a, p. 176.

de las organizaciones (con independencia de su carácter público o privado) es, precisamente, la inteligencia humana.

5. BIBLIOGRAFÍA

ABBOTT, Ryan: "I think, therefore I invent: creative computers and the future of patent law", *Boston College Law Review* (2016) Vol. 57, pp. 1079-1126.

ABBOTT, Ryan: "Everything is obvious", *66 UCLA Law Review (2019)*, 2, pp. 3-52.

BENTLY, Lionel, SHERMAN, Brad, GANGJEE, Dev, JOHNSON, Phillip: *Intellectual Property Law*, 5ª ed., Oxford University Press, 2018.

CALO, Ryan: "Artificial Intelligence Policy: a primer and roadmap", *University of Bologna Law Review* (2018) *3*(2), pp. 180-218.

CASADO CERVIÑO, Alberto: "Globalización y propiedad industrial: el escenario emergente", *Comunicaciones en Propiedad Industrial y Derecho de la Competencia* (2008), nº 49, pp. 63-77.

CERRILLO I MARTÍNEZ, Agustí: "¿Son fiables las decisiones de las Administraciones públicas adoptadas por algoritmos? *European Review of Digital Administration and Law- Erdal* (2020), Vol. 1, issue 1-2, pp. 18-36.

CERRILLO i MARTÍNEZ, Agustín: "El derecho para una inteligencia artificial centrada en el ser humano y al servicio de las instituciones", *Revista de Internet, Derecho y Política* (2020), nº 30, pp. 1-5.

CIERCO SIERRA, C: "El principio de precaución: reflexiones sobre su contenido y alcance en los Derechos comunitario y español", *Revista de Administración Pública* (2004), núm. 163, pp. 73-126.

EUROPEAN COMMISSION, DIRECTORATE-GENERAL FOR COMMUNICATIONS NETWORKS, CONTENT AND TECHNOLOGY, HARTMANN, Christian, ALLAN, Jacqueline, et al.: *Trends and developments in artificial intelligence – Challenges to the intellectual property rights framework – Final report*, Publications Office of the European Union, 2020.

EUROPEAN PATENT OFFICE, *Guidelines for examination*. Ed. 2024.

FERNÁNDEZ NÓVOA; Carlos, OTERO LASTRES, José Manuel; BOTANA AGRA, Manuel: *Manual de la propiedad industrial*, 3ª ed., 2017. Marcial Pons; Madrid.

GALLEGO SÁNCHEZ, Esperanza: "La patentabilidad de la inteligencia artificial. La compatibilidad con otros mecanismos de protección", *La Ley Mercantil (2019)*, nº 59.

GIL CELEDONIO, José Antonio: "Digitalización, inteligencia artificial y sus impactos jurídicos: una aproximación al caso de la propiedad industrial" en CANDELARIO MACÍAS, María Isabel (dir.): *La propiedad industrial en tiempos del COVID-19*. Tirant Lo Blanch, 2022, pp. 17-48.

GIL CELEDONIO, José Antonio: "La Inteligencia artificial aplicada en el sector público: el caso de las Oficinas de Propiedad Industrial y, en particular, el de la Oficina Española de Patentes y Marcas, O.A.", *Comunicaciones en Propiedad Industrial y Derecho de la Competencia* (2023), nº 99, pp. 71-95.

ÍÑIGUEZ, Pilar: "Intellectual Property Rights, Artificial Intelligence and Big Data: future perspectives", *Actas de Derecho Industrial (*2021), *nº 41*, pp. 109-132.

KERBER, Wolfgang: "Digital revolution, institutional coevolution and legal innovations", *European Business Law Review* 34 (2023), nº 6, pp. 993-1016.

KOULU, Riikka, HIRVONEN, Hanne, SANKARI, Suvi, HEIKKINEN, Tatjaan: "Artificial Intelligence and the Law: Can and Should We Regulate AI Systems?", en: *Edward Elgar Handbook on Law and Technology*, Edward Elgar, 2023, pp. 427-449.

KRAKUP, Troels, HORST, Maja: "European artificial intelligence policy as digital single market making", *Big Data and Society* (2023), January-June, 2023, 1-14

MAROÑO GARGALLO, María del Mar: "El concepto de inventor en el derecho de patentes y los sistemas de Inteligencia Artificial", *Cuadernos de Derecho Transnacional* (2020), vol. 12, nº 2, pp. 510-526.

MASSAGUER FUENTES, José: "La propiedad industrial: balance y perspectivas", *Actas de Derecho Industrial y derecho de la competencia* (1998), pp. 93-126.

MCLAUGHIN, Michael: "Computer-generated inventions", *Social Science Research Network Electronic Journal* (2018) nº 15, pp. 1-32.

PRESNO LINERA, Miguel Ángel: "Teoría General de los Derechos Fundamentales e Inteligencia Artificial: una aproximación", *Revista Jurídica de Asturias* (2022) nº 45, pp. 48-57.

PONCE SOLÉ. Juli: "Inteligencia Artificial, decisiones administrativas discrecionales totalmente automatizadas y alcance del control judicial: ¿indiferencia, insuficiencia o deferencia? *Revista de Derecho Público: Teoría y Método* (2024) Vol. 9, pp. 171-2020.

PONCE SOLÉ, Juli: "Límites jurídicos de la toma de decisiones discrecionales automatizadas mediante Inteligencia Artificial: Racionalidad, Sabiduría y necesaria reserva jurídica de humanidad en el ámbito digital", *Revista General de Derecho Administrativo* (2024), núm. 66.

PILA, Justine, TORREMANS, Paul: *European Intellectual Property Law,* 2nd ed., Oxford University Press, 2019.

SÁNCHEZ BARRILAO, Juan Francisco: "Inteligencia Artificial y fuentes del derecho", *Revista española de Derecho Constitucional Europeo* (2023), núm. 39.

SÁNCHEZ GARCÍA, Luz: "Propuesta de una nueva categoría de inventor para el sistema de patentes español. El inventor artificial", *Comunicaciones en Propiedad Industrial y derecho de la competencia* (2019), nº 86, pp. 77-90.

VAN CAENEGEM, William: *Intellectual Property Law and innovation*. Cambridge University Press, 2007.

VANHERPE, Josephine: "AI and IP- Great Expectations", en: DE BRUYNE, Jan, VANLEENHOVE, Cedric (eds.): *Artificial Intelligence and the Law, 2nd ed*. KU Leuven Centre for IT&IP Law Series, 2023, pp. 233-267.

VILLAGRASA, Óscar: "La inteligencia artificial del sector público: desarrollo y regulación de la actuación administrativa inteligente en la cuarta revolución industrial", *Revista de Internet, Derecho y Política* (2020), nº 30, pp. 1-14.

Capítulo 18

INTELIGENCIA ARTIFICIAL Y SERVICIOS PÚBLICOS: UNA VUELTA AL ORIGEN PARA CONSTRUIR EL FUTURO

Daniel Valls Broco
Investigador en formación
Facultat de Dret, Universitat de Barcelona

RESUMEN: Según un informe[1] hecho por el FMI en enero de este mismo año, **casi un 40% del empleo mundial se encuentra afectado por la Inteligencia Artificial;** en economías «avanzadas» esta cifra crece prácticamente hasta el 60%. En España se movilizaron 600 millones de euros de inversión en IA en el año 2020 y este 2024 se ha aprobado un plan de gasto de 1.500 millones de euros para los próximos 2 años. Estos son algunos datos que explican la necesidad y la importancia de poner el foco sobre la utilización de la IA.

Precisamente por ello analizaremos los **antecedentes** de esta tecnología; sus **riesgos**; su crecimiento *a ritmo de Moore*, la alargada sombra del progreso y la importancia de centrarnos en la lucha contra la *brecha digital* y la pobreza; y, visto el *engranaje* conceptual necesario, nos adentraremos en el complejo entramado de sujetos que conforman su puesta en escena en los servicios públicos, para acabar volviendo a la concepción originaria de estos, y, con el ciudadano como *eje*, poner de relieve sus reclamos históricos y las **oportunidades** de presente y de futuro que tiene ante sí la Administración para conseguir un **cambio de paradigma en la prestación de los servicios públicos**.

PALABRAS CLAVE: Inteligencia Artificial, Administración Pública, Servicios Públicos, Brecha Digital y Pobreza

1 CAZZANIGA, Mauro; JAUMOTTE, Florence; LI, Longji; MELINA, Giovanni; J. PANTON, Augustus; PIZZINELI, Carlo; ROCKALL, Emma; M. TAVARES, Marina: Gen-Al: Artificial Intelligence and the Future of Work. *IMF Staff Discussion Notes, Research Department* (2024) núm. 1, pp. 1-42.

1. INTRODUCCIÓN

La Inteligencia Artificial, desde el punto de vista normativo, es: «**un conjunto de tecnologías en rápida evolución** que contribuye a **generar beneficios económicos, medioambientales y sociales** muy diversos **en todos los sectores económicos y las actividades sociales**»[2]. Únicamente de ella puede deducirse a simple vista la vasta y amplia gama de contenidos afectos y/o relacionados con la IA y la dificultad de constreñirla en un reducido conjunto de palabras. El nuevo Reglamento sobre IA la categoriza como un «conjunto de tecnologías», lo que podemos intuir que busca simplificar al máximo la *vertiente tecnológica*; que la IA está en «rápida evolución», a lo que yo redefiniría como **incesante y vertiginosa** (en el buen y mal sentido de la palabra) **evolución**; que «**contribuye a generar beneficios**» (inapelable, no requiere de mayor análisis); y finalmente precisa que dichos beneficios son **principalmente** «**económicos, medioambientales y sociales**» (siendo algunos de los principales **objetivos de desarrollo de la Unión**), además de que son «muy diversos en todos los sectores», poniendo de manifiesto la **transversalidad de la IA**.

A nivel doctrinal, la IA es un tema de rigurosa actualidad (aunque con más de 2 siglos de historia) sobre el que se han vertido *ríos y ríos de tinta jurídica* (y más allá de ella), o, nunca mejor dicho, *miles de bits en nuestras pantallas*, como innovación transversal que **ha revolucionado significativamente la vida de las personas**; pasando de ser objeto de debate, con fervientes amantes e implacables detractores, a objeto de regulación de una de las cuestiones centrales que, irremediablemente, **rigen y regirán el Derecho del siglo XXI**, influyendo en ámbitos como el de los **servicios públicos**. No obstante, el origen y configuración de esta data de mucho antes, y su presencia en nuestras vidas también, así que viajaremos hasta su génesis.

2. UNA APROXIMACIÓN HISTÓRICA: DESDE EL TEST DE TURING Y EL VERANO DEL 55 HASTA LA *DEEP BLUE* Y LA *FRUSTRACIÓN DE KASPAROV*

Para remontarnos a los orígenes de la IA debemos situarnos prácticamente dos siglos atrás, en Lincolnshire (Inglaterra), donde *George Boole* afirmó, en 1854, que «la lógica podría sistematizarse de idéntica forma en que se resuelve un sistema de ecuaciones», la primera *chispa* conceptual que acabaría derivando en el nacimiento de la IA. A principios del siglo XX, en el Teatro Nacional de Praga, en 1921, *Karel Apek* acuñaría por primera vez el concepto de «robot» en la obra de teatro R.U.R. En 1936 volveríamos a Inglaterra, pero esta vez a Londres, donde *Alan Turing* publicaría un artículo en el que definiría el concepto de **algo-**

2 Definición empleada en el Considerando 4 del nuevo Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial.

ritmo, sentando así las bases de la informática. Posteriormente, en la Berlín de 1941 *Komrad Zuse* sería el artífice de Z3, la primera computadora programable y enteramente automática. Sólo 1 año después, en Estados Unidos, Isaac Asimov publicaría su relato *Runaround* (o *Círculo* Vicioso), en el que residían las leyes de la robótica. Casi 10 años después, en 1950, el británico *Alan Turing* publicaría en la revista *Mind* su famoso ensayo *Computing Machinery and Intelligence*, en el que presentaría lo que acabó conociéndose como el *test de Turing*[3].

Las *piezas* empezaban a disponerse sobre el *tablero*, y en una Estados Unidos repleta de figuras ilustres, su evolución empezó a acelerar. El verano de 1955 pasó a la historia porque un grupo de 4 científicos solicitaron financiación para un proyecto de investigación al *Dartmouth College* de New Hampshire, basándose en «la conjetura de que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tanta precisión que se puede hacer que una máquina lo simule» para «averiguar cómo hacer que las máquinas utilicen el lenguaje, formen abstracciones y conceptos, resuelvan tipos de problemas ahora reservados a los humanos y se mejoren a sí mismas». Ese «grupo cuidadosamente seleccionado de científicos» (como se autodefinieron) eran *Claude Shannon*, *Marvin Minsky*, *Nathaniel Rochester* y *John McCarthy*, **siendo este último, en 1956, quién organizó la conferencia de Dartmouth y dio nombre a la disciplina: IA[4].** 2 años después, en 1958, Frank Rosenblatt diseñó la primera red neuronal artificial: Perceptrón. Pero no fue hasta 1969 cuando Marvin Minsky y Seymour Papert publicaron *Perceptrons*, analizando las redes neuronales artificiales.

Una vez sentadas las bases de la disciplina se produjeron hitos históricos alrededor del mundo: en 1979 el coche ***Stanford*** sería uno de los primeros con **función autónoma**[5]; en 1980 se fabrica el ***Wabot-2***, el primer robot en realizar algunas **acciones humanas**[6]; en 1996 la supercomputadora ***Deep Blue***, creada por IBM, sería capaz de vencer al campeón del mundo de ajedrez *Gary Kaspárov*[7]; en 2012 *Google* creó una red neuronal informática capaz de aprender a identificar gatos, caras y cuerpos, a partir de los datos obtenidos por *YouTube* utilizando un sistema de *deep learning*; en 2014 llegó ***Eugene***, el programa de ordenador que, actuando como

3 La primera evaluación de comunicación verbal hombre-máquina, que pondría a prueba la capacidad de las máquinas para interactuar como si fuesen humanos.

4 Según el historiador *Jonnie Penn* el nombre de la disciplina fue objeto de discusión, planteándose diversas alternativas. *John McCarthy*, fruto de la ambiciosa visión que tenía de su investigación, la denominó IA, pretendiendo llamar la atención de todo el que lo escuchara, incluyendo posibles interesados en su financiación, lo que sin duda logró. Paralelamente, *Shannon* y otros investigadores utilizaron otros nombres, ya que no comulgaban con lo que significaba tildarla de *Inteligencia* Artificial.

5 *Stanford* cruzó con éxito una habitación llena de sillas sin ser conducido.

6 *Wabot-2* se comunicó con otra persona, leyó una partitura y tocó ciertas melodías en un órgano simple.

7 *Gary Kaspárov* fue desconcertado por los movimientos de la IA. Él mismo confesó que tuvo problemas para dormir por uno de sus movimientos. Su frustración fue tal que acusó a IBM de actuar fraudulentamente, programándola para provocar gran presión psicológica sobre él, y pidió que abrieran los *logs* (archivos que registran eventos específicos dentro de un sistema); *IBM* no accedió. *Kaspárov* quiso revancha, *IBM* no aceptó y el superordenador fue desmantelado.

si fuera un chico de 13 años, **consiguió superar el Test de Turing formulado más de medio siglo antes**; y en 2015, *AlphaGo* se convirtió en la primera IA capaz de ganar a un jugador profesional de Go sin utilizar *piedras de hándicap*[8] en un tablero de 19x19.

Paralelamente a ello, la IA se implementó en la vida de las personas, teniendo su muestra más ejemplificativa en los **algoritmos de los teléfonos móviles**, que obtienen de nuestros datos y nuestro uso, información que recopilan y utilizan para mostrarnos contenido publicitario, entre otras finalidades, en base a lo que cada uno consienta (dentro de lo que buenamente le dejen, según el tipo de *cookies* y ajustes existentes). Asimismo, esta tecnología generó, genera y generará riesgos, cuestión sobre la que arrojaremos luz a continuación.

3. UNA CUESTIÓN DUAL: A LA LUZ DE LOS BENEFICIOS DE LA IA, LOS TRANSLÚCIDOS RIESGOS ASUMIDOS

Más allá de la multiplicidad de usos de la IA y sus indiscutidos beneficios, volviendo a poner el foco sobre el Reglamento Europeo y su definición, esta enfatiza sus beneficios, pero no ha dedicado ni un atisbo a la sombra que la acecha: **el riesgo**, cuestión menos extraña aún si cabe para el ámbito de los servicios públicos, pues es frecuente el uso de esta terminología, tanto en su gestión, como en su prevención.

El riesgo nace en contraposición al peligro. El peligro es de origen natural, mientras que el riesgo radica en lo generado artificialmente, por el uso de la técnica, de la tecnología. Por consiguiente, la IA, como conjunto de tecnologías, conlleva riesgos. En realidad, el Reglamento, aunque no en su definición, sí contempla la existencia de riesgos y los clasifica en diversos tipos, condicionando su entrada en vigor a los citados riesgos. Estos inciden sobre derechos fundamentales, que ya se alertaban por estudiosos de la materia.

Por un lado, *Cerrillo i Martínez*[9], nos advirtió del «potencial transformador de la IA», y de «la necesidad de que su uso preserve adecuadamente valores esenciales para nuestra sociedad como la dignidad, la autonomía y la autodeterminación de las personas o principios como la no discriminación, el respeto de la legalidad, la transparencia y la comprensibilidad en los procesos de toma de decisiones». A ello añadió que es necesario un marco normativo y ético adecuado y señaló que los trabajos de investigación publicados eran insuficientes para dar respuesta a los retos de la IA. En efecto, no se equivocaba.

8 Piedras utilizadas en el tablero para que jugadores menos experimentados o de menor nivel puedan disputar partidas contra otros más experimentados o mejores con condiciones más favorables de inicio.

9 CERRILLO i MARTÍNEZ, Agustí: "El derecho para una inteligencia artificial centrada en el ser humano y al servicio de las instituciones", *IDP. Revista de Internet, Derecho y Política* (2020), núm. 30, pp. 1-6

Por otro lado, *Capdeferro Villagrasa*[10] apuntó que el uso de la IA en España no sólo era «desigual», sino que era «bastante desconocido» y añadió que sería recomendable contar con un «catálogo exhaustivo de todas las aplicaciones informáticas que se empleaban en el uso de la inteligencia artificial». También advirtió que el empleo de algoritmos puede entrar en tensión con principios como la seguridad jurídica, la transparencia o la obligación de motivación, o con derechos fundamentales como la igualdad, la intimidad o la protección de datos personales.

Adicionalmente, *Miró-Llinares*[11] distinguió dos posiciones diferenciadas frente a la IA: «los optimistas y los pesimistas». Los optimistas contemplan que su uso puede permitir mejorar la objetividad y la eficacia de la prevención y persecución del crimen. Los pesimistas alertan sobre su posible efecto discriminatorio, o sobre una hipotética falta de control ético o social en su uso e, incluso, de los peligros de construir unos Estados basados en el control y la vigilancia permanentes por medio de estas herramientas. Sobre esta base el autor propone efectuar una «**lectura realista, crítica y empíricamente informada del uso de los algoritmos y los datos masivos**».

Indudablemente la IA ha generado riesgos y con este tipo de *lectura* encontraremos ejemplos concretos. Antes analizaremos *a Moore*, el progreso y su sombra: la *brecha digital* y la desigualdad generada por la pobreza.

3.1. *Al ritmo de Moore*, la *Muerte Silenciosa* de la población sin recursos. «*Ideas Potentes*» en la lucha contra la *Brecha Digital* y la Pobreza

Los últimos años se encuentran impregnados por un crecimiento exponencial de innumerables avances que han contribuido a la mejora y al desarrollo de la sociedad, entre ellos, la IA. Para entenderlo es bastante ilustrativa la *ley de Moore*[12], que predijo la velocidad del cómputo del desarrollo de la tecnología, provocando una importante reducción de costes en su fabricación. Su impacto fue tal que no sólo se mantuvo desde que se formuló (1965), sino que fue fuente de inspiración para lo que vino después.

No obstante, el propio *Moore* en 2007 alertó en una entrevista a *The Inquirer*: «Mi ley dejará de cumplirse dentro de 10 o 15 años». En la actualidad, algunos afirman que la "ley"

10 CAPDEFERRO VILLAGRASA, Oscar: "La inteligencia artificial del sector público: desarrollo y regulación de la actuación administrativa inteligente en la cuarta revolución industrial". *IDP. Revista de Internet, Derecho y Política* (2020), núm. 30, pp. 1-14.

11 MIRÓ-LLINARES, Fernando: "Predictive Policing: Utopia or Dystopia? On attitudes: towards the use of big data algorithms for law enforcement", *IDP. Revista de Internet, Derecho y Política* (2020), núm. 30, pp. 1-18.

12 *Gordon Moore*, en sentido estricto y en términos científicos no creó una ley, sino que formuló una predicción en su artículo *Cramming more components onto integrated circuits* para la revista *Electronics*, partiendo del análisis de un gráfico del número de transistores anual que explica que se duplique el número de transistores en un microprocesador cada año. En 1975 su predicción pasó a ser cada 2 años.

murió hace años, con la reducción del ritmo de fabricación en los procesadores *Intel*. En respuesta, *Jim Keller* (director de la ingeniería de silicio de *Intel*) afirmó: «llevo oyendo hablar del final de la Ley de Moore desde que empezó mi carrera» y que: «al cabo de un tiempo decidí dejar de preocuparme por eso», además de asegurar que *Intel* tenía trazada la hoja de ruta para los próximos 10 años. En este contexto, *Pat Gelsinger*, *CEO* de *Intel*, reconoció que la IA es clave para recuperar su empresa. Quizás sea la llave para volver al tan ansiado *ritmo de Moore*, o al menos, a uno parecido.

Ahora bien, en palabras de *George Olah*: «Si no crees que los modelos de IA vayan a ser muy capaces, entonces probablemente no crees que vayan a ser demasiado peligrosos». Como entraña la afirmación de *Olah* sabemos que son capaces, también peligrosos, por lo que debemos plantearnos: ¿Cuál es el coste de recuperar el *ritmo de Moore*? ¿La IA es la respuesta? ¿El progreso debe ser nuestra única preocupación?

En 2020, *David Rotman* en su artículo «Ha llegado la hora de que cunda el pánico por el fin de la ley de Moore» publicado en la *MIT Technology Review*, alertaba que «el coste de las fábricas de chips más avanzados aumenta en alrededor del 13% anual, y se espera que roce los 15.000 millones de euros para 2022». Y añadió que *Neil Thompson*: «muestra un amplio margen para mejorar el rendimiento computacional», lo que «parece buena noticia para el progreso continuo, pero a Thompson le preocupa que eso también señale el declive de los ordenadores como tecnología de propósito general (...) lo que supondrá una ventaja para aquellos con suficiente dinero y recursos». Finalmente el autor plantea las ventajas de un *plan Marshall* para los *chips*; apunta que *Thompson* llama a la inversión pública, pero pone en entredicho las garantías de tales inversiones; y reflexiona sobre que **«es casi seguro que siempre vamos a querer más potencia informática».**

Paralelamente al progreso, habida cuenta del uso masificado de la IA, no debe pasar inadvertida la alargada sombra que acecha al mundo: la pobreza, históricamente asociada a la escasez de dinero y recursos. Actualmente, con la irrupción de nuevas tecnologías, sus riesgos han generado una nueva forma: la pobreza digital. *Margaret Boden*, una influyente científica cognitiva, en una entrevista de la BBC en 2014, advirtió sobre que no bastaría con tener ordenadores potentes, sino que el campo de la IA también necesitaría «**ideas potentes**». En búsqueda de *ideas potentes*, con las cifras de *Rotman*, y siguiendo los temores de *Thompson*, tenemos una gran responsabilidad: combatir la *brecha digital* y la pobreza en todas sus formas.

La comunicación, el ocio y muchos otros aspectos fundamentales para la población se cimentan en la tecnología. Paralelamente ha irrumpido en el discurso de los medios y de los políticos. También ha llegado a la educación, con herramientas como el *Campus Virtual*, los *fórums* o materiales únicamente accesibles en plataformas digitales. En este contexto, acrecentándose la masificación de la IA y la *brecha digital*, no está todo perdido para aquellos en situación de pobreza, o de pobreza extrema. Algunas entidades han creado movimientos y planes desarrollados para combatir la pobreza digital. En esta dirección, los Estados juegan un papel fundamental, especialmente en la prestación de servicios públicos.

4. IA Y SERVICIOS PÚBLICOS: UNA OPORTUNIDAD ÚNICA PARA OPTIMIZAR UN COMPLEJO ENTRAMADO CON SUJETOS CONDENADOS A ENTENDERSE

Los servicios públicos se encuentran ante un nuevo escenario. *Velasco Rico*[13] ya alertó sobre la imperiosa necesidad de que el servicio fuera «**personalizado y diferenciado para cada uno de los ciudadanos, en base a sus pretensiones y necesidades**, además, por supuesto, de por los **intereses públicos** que motivan la prestación de estos», y añadió que «debe dejar de ser únicamente un sujeto pasivo que recibe el servicio público, para pasar a ser parte activa de este», siendo esta la aspiración y el fin último.

4.1. La Administración y un *software* que *actualizar*: Confianza, Eficiencia e Interacción.

La Administración no sólo tiene una responsabilidad y obligación para con el ciudadano, sino para sí misma y para con la sociedad. La IA no debe *cegar* a la Administración en búsqueda de mejoras técnicas únicamente, sino que debe recordar los reclamos históricos de la ciudadanía. Su ejemplo más reciente son las (ya habituales) quejas sobre la red de servicios ferroviarios en España, así como en el estado de sus infraestructuras; si bien existen otras deficiencias notables, como en la gestión del servicio del agua y de electricidad. La Administración necesita actualizarse cual *software* con nuevas versiones, pero también dejar de *parchear* problemas que acabarán llevando los servicios hasta un *punto de no retorno.* Para ello deberá centrarse en 3 pilares: **confianza, eficiencia e interacción.** La respuesta reside en el ciudadano.

4.2. El Ciudadano como eje de una *actualización* necesaria. Sin *excusas* ante reclamos históricos más allá de una IA Ética, Transparente y Humanitaria.

La figura del Ciudadano es el eje central, dando sentido a la categorización del servicio como público. Quizás en el pasado bastara con un servicio generalizado, pues resultaba inconcebible la personalización y adaptadabilidad de este respondiendo, a su vez, al interés general. Los tiempos han cambiado, también la perspectiva. El grado de ambición en la prestación del servicio debe alinearse con las expectativas y exigencias de la *sociedad de la inmediatez*, recuperando *su bien más preciado*: **la confianza**.

La IA debe ser una herramienta que mejore servicios, no un obstáculo en la recepción final del usuario. Su millonaria inversión debe acompañarse de *ideas potentes*

13 VELASCO RICO, Clara Isabel: "Personalizacion, proactividad e inteligencia artificial ¿Un nuevo paradigma para la prestación electrónica de servicios públicos?". *IDP. Revista de Internet, Derecho y Política* (2020), núm. 30, págs. 1-16.

y de alternativas, para aquellos que, por imposibilidad (pobreza), no puedan ser destinatario final de la *actualización* en los servicios. Cabe mantener formas tradicionales y/o paralelas, en la figura del funcionario, acompañando y asegurándose de la efectividad del servicio; extremando todavía más las precauciones con la consolidación de la IA.

No es sólo una *actualización* necesaria, sino la *excusa* perfecta para que la Administración se deje de *excusas* en la garantía de ofrecer servicios *de minimis* en términos de calidad. La oportunidad para responder a *la llamada del ciudadano* y acabar con la desconfianza, el recelo y el pesimismo. Para ello, la **eficiencia** es la respuesta a la desconexión instituciones-ciudadano. Acabar con situaciones como: números de teléfono asignados a la Administración en desuso; la redirección a otras instituciones o a terceros para eludir responsabilidades, ahora también utilizando *bots* o *chat bots*; la falta de atención a casos más extraños o residuales, sin respuesta ante *bots* o *chat bots*; o la, en ocasiones, excesiva limitación de atención directa al ciudadano.

Paralelamente, la Administración muestra un alto nivel de exigencia hacia el ciudadano. Último ejemplo de ello son las problemáticas y reclamos generados con posterioridad a las concesiones del IMV, afectando los hogares de las personas más vulnerables, que también son algunos de los principales usuarios de servicios públicos. Ello sucedió en uno de los peores momentos posibles, una pandemia mundial, que dejó heridas de muerte las ya maltrechas economías de esos ciudadanos, a la par que su confianza en el sistema. No obstante, no ahondaré más en cuestiones específicas, puesto que, si bien podría dar una larga lista, *lanzar dardos al vacío* no es el objetivo, sino *dar en la diana*, señalando situaciones que deben cesar para crear un ambiente propicio a la interacción con el ciudadano.

Este caldo de cultivo imposibilita el éxito de los servicios públicos en una concepción personalizada, condenándolo a un injusto ostracismo. La Administración debe *tomar cartas en el asunto*, evitando que una situación ya desfavorable acabe por convertirse en una fragmentación (o incluso quiebre) de la relación Administración-Ciudadano. Para ello debe comprender y potenciar al máximo el uso del algoritmo, y aquí es donde el programador y/o proveedor juega un papel fundamental.

4.3. El Proveedor - Programador: las Consecuencias del Secreto Profesional y las Alternativas a la cesión de la Llave: la *Fuente* del *Código*

El proveedor y/o programador es la *llave* para la implementación de la IA en los servicios públicos. Su figura ha generado un intenso debate sobre si debería facilitar sus secretos a la Administración: el *Código Fuente*. Los defensores de la cesión esgrimen que la IA es información pública, al formar parte de un servicio público. Dicha exigencia parte de la premisa del derecho ciudadano a obtener respuestas motivadas; habiéndose limitado la Administración en estos años a contestar que la respuesta es la facilitada por el algoritmo, sin mayor explicación. Ante los Tribunales añadió que no podría motivarse de otra forma, al desconocer el funcionamiento del algoritmo y no disponer del *Código Fuente*.

Ahora bien ¿sería justo exigir al programador la cesión del *Código Fuente*?

Lo cierto es que entregarlo equivaldría a quitarle todo elemento de valor, este es el corazón y el cerebro del algoritmo, y en efecto, contiene sus secretos y explica su funcionamiento. Su pérdida podría tener consecuencias irreversibles. Además, no es ético, y menos, seguro, despojarlo de este, siendo la figura más capacitada para tomar decisiones, tanto en pro del beneficio público, como para su posible interés privado, ya habituado a tomar medidas de protección específicas como mayor interesado en proteger a toda costa el *Código Fuente*. Por todo ello es su único y mejor guardián.

En términos de ciberseguridad son frecuentes los ataques sufridos por Administraciones Públicas, especialmente en Europa. Por ende, el hecho de que el programador ponga el *Código Fuente* a disposición de la Administración aumentaría el riesgo de ser atacados, y con ello, de desmantelar o alterar el funcionamiento del algoritmo. Incluso es una cuestión numérica: cuantas más personas conozcan la *llave*, menos seguro será, ya que las contraseñas son y deben ser privadas. Poniéndonos catastrofistas, cualquiera con la información y los conocimientos necesarios podría *sacarle partido*.

Entonces ¿qué hacemos?

Alternativamente a su cesión sería más prudente que el programador y/o proveedor explique (con todo lujo de detalles) el funcionamiento del algoritmo y sus decisiones y respuestas. El legislador podría establecer exigencias y límites. Además, al tratarse de una relación contractual pueden establecerse cláusulas (incluso previstas o limitadas por el legislador). Así se tendría pleno respeto por el secreto profesional, y pasaría a exigirse *la Fuente del Código*, no el *Código Fuente*.

Cuestión distinta son las responsabilidades generadas por malas prácticas de la Administración o por deficiencias o fallos del algoritmo, sea por instrucciones mal dadas por la Administración al programador en su confección; sea porque el programador haya cometido algún error en su configuración; o, sea el propio algoritmo quien empiece a fallar. Afortunadamente, estas serán (o al menos así debería ser) cuestiones residuales de las que deberá hacerse cargo la Administración y/o el programador según la situación, lo establecido en el contrato y la normativa vigente en ese momento. Adicionalmente, podría valorarse la habilitación expresa del legislador a los titulares de los Juzgados y Tribunales para poder exigir la cesión del *Código Fuente* en casos específicos y/o de especial complejidad que así lo requieran, para descubrir la raíz del problema y determinar responsabilidades. En este contexto, las Agencias previstas por el nuevo Reglamento de IA podrían desempeñar un papel fundamental.

5. CONCLUSIONES

En definitiva, cómo ya avanzó *Boden*, la IA debe ir acompañada de ***ideas potentes***, y con ellas crear **soluciones**. Analizados los precedentes y los riesgos en el uso de estas tecnologías nos encontramos ante un nuevo horizonte. Una **oportunidad** única **para que la Administración Pública no sólo actualice y optimice la prestación de servicios**

públicos, sino para que redefina y revolucione por completo su realidad, ayudando a combatir la *brecha digital* y dando respuesta al **interés general**.

Ello requiere de **personalización** y **proactividad**: garantizando la **atención directa** y la **proximidad** con el ciudadano; respondiendo a sus **intereses** y **necesidades**; y recuperando la tan ansiada **confianza** e **interacción** con el usuario, para evitar que se repita la *frustración de Kaspárov* y su desconfianza en la IA, pero esta vez de forma generalizada.

Las inversiones millonarias de dinero que conlleva deben ir acompañadas de **mejoras en los reclamos y quejas históricas que la población ha transmitido a las instituciones durante estos años.** La IA tiene que ser la *punta de lanza* en la **lucha contra la pobreza**, contribuyendo a la **igualdad**, también digital, y **generando recursos** que recuperen el verdadero sentido originario de las instituciones públicas, nunca mejor dicho, **al servicio del ciudadano.**

6. BIBLIOGRAFÍA

CAPDEFERRO VILLAGRASA, Oscar: "La inteligencia artificial del sector público: desarrollo y regulación de la actuación administrativa inteligente en la cuarta revolución industrial". *IDP. Revista de Internet, Derecho y Política* (2020), núm. 30, pp. 1-14.

CAZZANIGA, Mauro; JAUMOTTE, Florence; LI, Longji; MELINA, Giovanni; PANTON, Augustus J.; PIZZINELI, Carlo; ROCKALL, Emma; TAVARES, Marina M.: Gen-Al: Artificial Intelligence and the Future of Work. *IMF Staff Discussion Notes, Research Department* (2024) núm. 1, pp. 1-42.

CERRILLO i MARTÍNEZ, Agustí: "Reptes i oportunitats de l'ús de la intel·ligència artificial a les administracions públiques", *Oikonomics* (2019), núm. 12, pp. 1-7.

CERRILLO i MARTÍNEZ, Agustí: "El derecho para una inteligencia artificial centrada en el ser humano y al servicio de las instituciones", *IDP. Revista de Internet, Derecho y Política* (2020), núm. 30, pp. 1-6.

MCCARTHY, John; MINSKY LEE, Marvin; ROCHESTER, Nathaniel; SHANNON ELWOOD, Claude: "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence" (1955), pp. 1-13.

MIRÓ-LLINARES, Fernando: "Predictive Policing: Utopia or Dystopia? On attitudes: towards the use of big data algorithms for law enforcement", *IDP. Revista de Internet, Derecho y Política* (2020), núm. 30, pp. 1-18.

TURING, Alan Mathison: "Computing Machinery and Intelligence". *Mind* (1950) núm. 49, pp. 433-460.

VELASCO RICO, Clara Isabel: "Personalizacion, proactividad e inteligencia artificial ¿Un nuevo paradigma para la prestación electrónica de servicios públicos?". *IDP. Revista de Internet, Derecho y Política* (2020), núm. 30, pp. 1-16.

Capítulo 19

A REVIEW OF HIGH-RISK ARTIFICIAL INTELLIGENCE (AI) SYSTEMS THAT ASSESS SOCIAL SECURITY ELIGIBILITY

Mariah Brochado

Chair of Philosophy of Technology and Digital Rights (UFMG). Visiting Professor at the Leibniz-Institut für Medienforschung | Hans-Bredow-Institut. President of the Artificial Intelligence in Law of Minas Gerais' Bar commission.

Lucas Porto

PhD in Law (UFMG). Senior researcher at the Chair of Philosophy of Technology and Digital Rights (Philotech – UFMG). Socialenvironmental Scientist. Lawyer.

Amanda Mapa

PhD student in Law (UFMG). Master of Laws (UFOP). Research director at Brazil's Social Security Studies Institute. Lawyer.

ABSTRACT: From 2020, as debate on AI regulation by legislative bodies looms, social security agencies in France, the UK and Brazil have increasingly used automated and AI decision systems to assess eligibility for social security benefits. By 2024, AI legislation will take root, requiring a review of the regulatory legitimacy of federal agencies and their ongoing use of AI systems. The European Union's AI law bans general-purpose scoring systems and classifies systems used by public authorities to assess social security benefits as high-risk, affecting fundamental social rights. However, current practices in AI eligibility assessment have damaged citizens' social rights. For example, France's CAF reported that 100,000 citizens have been wrongly denied benefits since 2021, Brazil's INSS found an increase in benefit denials with automated systems, and the UK's AI system overestimated income, leading to inappropriate benefit denials. These practices illustrate the potential injustices that AI can cause and highlight the need for comprehensive regulatory frameworks. Cases from Australia and Brazil highlight the serious consequences of automated social protection systems, and show the need for fairness, transparency and human dignity in AI systems used in social protection.

KEYWORDS: Artificial Intelligence; Social Security; Automated Decision Systems; Regulatory Framework; Eligibility Assessment.

1. INTRODUCCIÓN

From 2020, while the debate on the regulation of artificial intelligence (AI) by legislative bodies is still on the horizon, social security agencies in countries such as France, the United Kingdom (UK) and Brazil start to increasingly use automated and AI decision systems to assess the eligibility and level of social security benefits. By 2024, AI legislation is beginning to take root, and not only the regulatory legitimacy of federal agencies in relation to AI systems, but also their ongoing use, will need to be reviewed.

The Artificial Intelligence Act (AIA), (Regulation (EU) 2024/1689, approved by the European Union (EU), prohibits "general purpose" scoring systems that could reduce individuals to a single social score that affects various aspects of their lives (Article 5, c, AIA). While the AIA classifies as "high risk" (Annex III(5)(a)) "AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for essential public assistance benefits and services, including healthcare services, as well as to grant, reduce, revoke, or reclaim such benefits and services" it does not fully cover current AI scoring practices that assess social security benefits.

Therefore, while it's clear that systems that evaluate and score natural persons' data to decide on social security benefits are classified as high risk, as they pose a significant risk (Article 6, 3, AIA) to the fundamental right to social security recognised by the Charter of Fundamental Rights of the European Union (Article 34, "the Charter"), there are current AI eligibility assessment practices that have damaged citizens' social rights. Acoording to Le Cuadernique du Net (2022), France's Caisse des Allocations Familiale (CAF) claims that around 100,000 citizens have been wrongly denied benefits since 2021. Brazil's Instituto Nacional de Seguridade Social (INSS, 2024) says that automated analysis increased from 17% to 23% between 2022 and 2023, and that rejected social security claims increased from 50% to 70% since the introduction of AI systems for benefit eligibility. And the United Kingdom (UK) uses an AI system that has been overestimating people's income. All these initiatives impact directly in citizens' lives as social security is tied to vulnerability criteria, such as unemployment and employment injuries and preventing citizens from starving or giving them bare minimum life conditions (Maxwell, 2021).

We expect to examine and contribute to the regulatory framework for AI through data and legal comparison methods, as one of the next main steps in AI implementation is to understand its legal categories (in this case, high risk and social scoring) through empirical data and already ongoing uses of AI systems, such as AI assessment of eligibility for social security benefits.

2. FRAMING THE SYSTEMS FOR RECOGNISING SOCIAL PROTECTION RIGHTS IN THE EUROPEAN AI REGULATION

Risks are inherent to work and life, to human beings themselves. Social protection against these risks, as a public policy, appears to be a way of holding together the social fabric, be-

cause it promotes a healthy society, without condemning to fate individuals who, for whatever reason – illness, old age, death, pregnancy – are unable to carry out an activity that guarantees their own subsistence and that of their family.

Although still in its infancy, Wagner Balera (2010) points out that the *Act for the Relief of the Poor*, passed in England in 1601, was the first model of a permanent social protection programme, in which the state would be responsible for providing assistance to the neediest in situations of illness, disability and unemployment.

After a slow and progressive evolution, the Treaty of Versailles, which sealed the peace at the end of the First World War, created the world's first specialised social relations body, the International Labour Organisation (ILO), in 1919. The treaty stated that "[...] the building of peace should be based on the cornerstone of social justice" (Balera, 2010, p. 72), highlighting the need for a social security programme, which was later addressed in conventions and recommendations.

ILO Convention 102/1952 established minimum standards for social security, containing detailed and complete provisions for the payment of specific benefits to cover certain social risks defined therein. From then on, the right to social protection, rooted in the concept of human dignity, became part of humanity's legal heritage

The fundamental right to social security is recognised by the Charter of Fundamental Rights of the European Union (Article 34, "the Charter"), and Regulation 883/2004 and its implementing Regulation 987/2009 are the main documents in the European Union dealing with social security.

However, while on the one hand there has been regulatory progress and a real improvement in social protection formulas, on the other hand this new cyber-law has not yet been fully implemented.

The basic premises of protection and the minimum social risks to be covered by the state were established; in other words, awareness of the indispensability of social protection was established globally. However, while the introduction of AI systems in this field can extend the range of protection, so that new subjects and new needs can be met in a timely manner, it can also exclude people who, despite their need for social protection, are unable to satisfy the virtualised process. People who, it should be emphasised, are in a moment of vulnerability and are looking for support from the state, and who may encounter technological obstacles to the realisation of their right, as will be demonstrated.

The recently published European regulation on AI is based on the differentiation of risk levels of AI systems and imposes interventions proportional to them. Thus, the regulation moves from an absolute ban on systems whose risks are unacceptable, to strict regulation of high-risk AI systems, and finally to almost deregulation of AI systems considered low or no risk, for which only transparency requirements and some incentives to adopt codes of conduct are established.

Thus, by prohibiting the use of unacceptable AI systems, mitigating the risks of applications of high concern and adopting a *laissez-faire* approach in areas of low concern, European

regulation seeks to combine its innovation and competitiveness agenda with ethical and safety objectives (Paul, 2023).

In a specific analysis of "high risk", which is the focus of most of the regulation, although Art. 3 of the Regulation provides some concepts, the Regulation doesn't address it comprehensively. Legal scholar work has been vocal about how the next steps should be towards

As for the *high-risk sphere*, its semiotic articulation and regulation serve to provide legal certainty and reduce potential liability risks for harms for those seeking to innovate and deploy AI systems which might violate individual rights. The Commission creates a space for "safe" and tightly checked innovation, including in the public sector, both through the narrow scoping of the high-risk sphere to a limited number of applications and through concrete instruments such as the conformity assessment and CE label (Paul, 2023).

Therefore, we present seven cases of current automated or AI systems that assess social security benefits.

Current social security benefits assessment[1]			
Country/ Region	**Incident**	**Description**	**Consequences**
Netherlands	Incident 101	Family allowance system falsely accused thousands of beneficiaries of fraud. Since 2013, 26,000 innocent families were wrongfully accused, forced to repay amounts, leading to financial ruin and psychological harm. The error was partly due to a discriminatory algorithm treating having a second nationality as a risk factor.	Parliamentary inquiry commission published the final report "Ongekend onrecht". Prime Minister Mark Rutte and his cabinet resigned in January 2021.

1 The spreadsheet presented has some data taken from the Artificial Intelligence Incident Database, duly referenced at the end, which is an initiative that collects and documents incidents related to the use of artificial intelligence. The platform aims to increase transparency and accountability around the development and use of AI by providing information on failures, ethical problems and other challenges associated with this technology.

Current social security benefits assessment[1]			
Country/ Region	**Incident**	**Description**	**Consequences**
United Kingdom	Incident 189	Benefit fraud detection algorithm disproportionately targeted people with disabilities. DWP used advanced AI to track possible fraud but refused to disclose the algorithm. Facing fresh legal action to reveal the algorithms used.	People with disabilities feel terrorized and anxious. Second legal letter sent by the same group on the subject, with the first going out before Christmas.
United Kingdom	Incident 611	Suspensions of Universal Credit benefits for many Bulgarian nationals. Potential nationality-based targeting for benefit fraud investigations led to poverty and homelessness. Equality impact assessment identified disproportionate flagging of marriages from Greece, Albania, Bulgaria, and Romania.	The UK government risks contempt of court unless it improves its response to transparency requests regarding the use of AI for vetting welfare claims, according to the information commissioner.
USA (Freep, 2023)	$20 million settlement	Class-action lawsuit filed by unemployment insurance claimants wrongfully accused of fraud by the MiDAS automated system. Claimants eligible for compensation if they received a fraud determination between Oct. 1, 2013, and Aug. 31, 2015, and had assets seized on or after March 9, 2015.	Settlement approved by the Michigan Court of Claims.

Current social security benefits assessment[1]			
Country/ Region	**Incident**	**Description**	**Consequences**
France (La Cimade, 2019)	Defender of Rights report	Digitization of social administrations increased inequalities in access to social rights, especially for the most vulnerable and foreigners. Reduction of personnel and physical service points. Digital divide exacerbated; many lack access to computers and smartphones.	Expanded digital control reflects xenophobic policies. Coordinated effort by collectives and associations needed to combat digital discrimination.
Canada (Offices of the Auditor General of Ontario, 2015).	Social Assistance Management System	In 2014, Ontario automated social assistance decisions using IBM Cúram software. Auditor General reported 1,132 errors related to eligibility and payment amounts, totaling $140 million CAD in errors. Total system costs reached $290 million CAD by the end of 2015.	System errors caused significant problems for beneficiaries, resulting in loss of trust and financial challenges.
Brazil (INSS, 2024).	"MEU INSS" Digital Project	Digital Governance Policy started in 2016, expanded in 2020 by the Digital Government Strategy. "MEU INSS" Digital Project launched in 2017 to reduce in-person requests. Audit verified significant increases in automatic denials, indicating risks of wrongful decisions.	Significant increase in automatic denials from 2021 to 2022. Risk of increased appeals and legal actions due to wrongful decisions.

Current social security benefits assessment[1]			
Country/ Region	**Incident**	**Description**	**Consequences**
Sweden (European Commission, 2024)	Digital Economy and Society Index	Sweden ranks fourth out of 27 EU member states in digitalization, excelling in connectivity and use of digital technologies. The advanced integration of digital technologies impacts AI system results in social protection positively.	High level of digitalization and integration of digital technologies results in better outcomes in the application of AI systems in social protection.
Australia (Services Australia, 2024)	Robodebt	Automated system within Centrelink to check reported income. Averaging income led to inaccurate assessments, resulting in debts totalling over A$1.7 billion against 430,000 people. Parliamentary committee found robodebt had a devastating impact on emotional and financial well-being.	Federal Court of Australia declared robodebt system illegal, and the government agreed to refund wrongly charged amounts.

The great concern and strict regulation of high-risk AI systems is obvious, especially considering the potential negative impact on a particularly sensitive area: fundamental rights. According to points 48 and 58 of the initial considerations of the European regulation, these rights include: the right to human dignity, information, non-discrimination, workers' rights, the rights of people with disabilities and, above all, access to certain essential services and benefits, to which the regulation pays particular attention:

Another area in which the use of AI systems deserves special consideration is the access to and enjoyment of certain essential private and public services and benefits necessary for people to fully participate in society or to improve one's standard of living. In particular, natural persons applying for or receiving essential public assistance benefits and services from public authorities namely healthcare services, social security benefits, social services providing protection in cases such as maternity, illness, industrial accidents, dependency or old age and loss of employment and social and housing assistance, are typically dependent on those benefits and services and in a vulnerable position in relation to the responsible authorities. If AI systems are used for determining whether such benefits and services should be granted, denied, reduced, revoked or reclaimed by authorities, including whether beneficiaries are

legitimately entitled to such benefits or services, those systems may have a significant impact on persons' livelihood and may infringe their fundamental rights, such as the right to social protection, non-discrimination, human dignity or an effective remedy and should therefore be classified as high-risk. Nonetheless, this Regulation should not hamper the development and use of innovative approaches in the public administration, which would stand to benefit from a wider use of compliant and safe AI systems, provided that those systems do not entail a high risk to legal and natural persons. In addition, AI systems used to evaluate the credit score or creditworthiness of natural persons should be classified as high-risk AI systems, since they determine those persons' access to financial resources or essential services such as housing, electricity, and telecommunication services. AI systems used for those purposes may lead to discrimination between persons or groups and may perpetuate historical patterns of discrimination, such as that based on racial or ethnic origins, gender, disabilities, age or sexual orientation, or may create new forms of discriminatory impacts. However, AI systems provided for by Union law for the purpose of detecting fraud in the offering of financial services and for prudential purposes to calculate credit institutions' and insurance undertakings' capital requirements should not be considered to be high-risk under this Regulation. Moreover, AI systems intended to be used for risk assessment and pricing in relation to natural persons for health and life insurance can also have a significant impact on persons' livelihood and if not duly designed, developed and used, can infringe their fundamental rights and can lead to serious consequences for people's life and health, including financial exclusion and discrimination. Finally, AI systems used to evaluate and classify emergency calls by natural persons or to dispatch or establish priority in the dispatching of emergency first response services, including by police, firefighters and medical aid, as well as of emergency healthcare patient triage systems, should also be classified as high-risk since they make decisions in very critical situations for the life and health of persons and their property. (paragraph 58, p. 55).

Using cognitive robotics and process automation, the Municipality of Trelleborg in Sweden began automating social assistance decisions in 2016, allowing the programme to process and grant subsidies for sickness, unemployment and tax exemptions through a robotic decision-making system. The result, according to the report on AI in public services in the European Union, was a significant improvement in waiting times for applications to be analysed. However, the report itself, while positive, also highlights at least some areas of concern: "Some observers expressed concerns about the risk of excluding some more vulnerable citizens when all processes are automated online, as this makes it more difficult to assess individual needs" (Misuraca, G., van Noordt, 2020, p.46). And again:

While the AI-system enabled various social welfare benefits decisions to be automated, many other processes of the Trelleborg municipality still operate as in a traditional bureaucratic system. There are still many paperbased processes within the organisation which could lead to double documentation and inefficient processes, as well as existing software with very poor interfaces and usability levels. Hence. Process Automation in general and AI systems in particular can strongly improve one specific government process, but the interoperability with other organisational processes should never be forgotten (Misuraca, G., van Noordt, 2020, p.46).

Although the result for Sweden is optimistic, it is necessary to consider the context in which the Swedish government and recipients of social protection operate in order to make comparisons. According to the latest Digital Economy and Society Index published by the European Commission, which has been monitoring member states in the digital field since 2014 and analysing their level of digitalisation and progress in this area, Sweden ranks fourth out of 27 member states and continues to excel in connectivity and the use of digital technologies (European Commission, 2024). "Digital technologies, both existing and emerging, are more widely used and integrated in the developed countries of the EU, which has a direct impact on the results obtained through the application of AI systems in the field of social protection" (European Commission, 2024).

However, the reality in Brazil is quite different. According to the United Nations Development Programme (UNDP) Human Development Report 2023-2024, the country ranks 89th out of 193 nations, indicating significant digital inequality (UNDP, 2024). According to the IBGE, 25 per cent of the Brazilian population does not have access to the Internet, which equates to around 54 million people (IBGE, 2024).

Despite this, the body responsible for social protection benefits and services, the INSS, has automated its services. Through its platform, access to social services and benefits from the Brazilian government. As a result, social protection is now only available via the internet and the vast majority of benefits are automatically analysed for eligibility (INSS, 2024).

Against this backdrop, the agency itself has demonstrated the increasing rejection of applications. While in 2018 there were 3,889,600 rejections for 5,123,777 grants, in 2022, the last year recorded in the report, there were 5,113,354 rejections for 5,212,631 grants (INSS, 2024). This is highly problematic given the need to care for the vulnerable and meet the specific needs of the most marginalised and excluded groups in an unequal society, such as low-income earners, the elderly, the disabled and the sick.

Australia also has a negative experience. The country has an agency (Centrelink) that is central to the delivery of government services and benefits to the Australian community (Services Australia, 2024). In detail, the agency is responsible not only for the social protection of pre-defined vulnerable groups (unemployment, old age, housing problems due to natural disasters or domestic violence, alimony), but also for the recovery of allegedly wrongly received payments (Services Australia, 2024).

In the Australian welfare system, social security entitlements are based on individual income, so recipients declare their income to the government every two weeks, and the government uses this information to calculate their eligibility and the amount of payment they are entitled to receive. Within the agency, Robodebt has been developed, an automated system to check that recipients have correctly reported their income (Services Australia, 2024).

Australia's automated system has proved problematic for incorrect income and social security payments.

"The system overlooked a critical issue: averaged income is a poor guide to actual income in any particular fortnight. Averaging obscures variations in a person's actual earnings from fortnight to fortnight, which are essential for an accurate assessment of their entitlements.

The Government nevertheless insisted on using averaging alone to claw back social security payments from a vast number of people. Since 2015, when robodebt was first established, the Government has raised debts totalling over A$1.7 billion against 430,000 people – mostly aged pensioners, disability support pensioners, financial and mental health counselling. In September 2020, a parliamentary committee found that robodebt had "an overwhelming and devastating impact" on many people's emotional and financial wellbeing and willingness to engage with and trust government services" (Maxwell, 2021, p. 8).

The situation was so alarming at the time, and had such serious consequences, that in 2020 a committee of people who had been unfairly harmed by the system was set up, and a website created, where they could report the devastating personal consequences and band together to stop the money being returned, which it was. In the end, the Federal Court of Australia declared the robodebt system illegal, which the government itself admitted, and agreed to refund the amounts wrongly charged (Maxwell, 2021).

A very similar situation occurred in the UK, where more than 200,000 people were wrongly charged for housing benefit. The automated system flagged several people as fraudulent, but was later shown to be flawed and biased against vulnerable claimants (Maxwell, 2021).

On that matter, Eubanks explores the impact of automated algorithms in social policy and includes real-life accounts of people who have suffered digital social exclusion. One of the stories involves a young woman who received medical benefits, food stamps, public transport and other services from the US government (Eubanks, 2018).

3. CONCLUSIONS

The appeal of AI systems for social protection lies in their potential to work on a larger scale, as the Swedish experience shows - speeding up eligibility checks allows more assessments to be carried out in a short period of time and reduces costs. However, this tactic has also been controversial, as illustrated by the different examples from different countries.

The case of Australia's Robodebt system highlights the danger of increasingly automating social protection through AI. A system that relied on average income as a measure of eligibility, which ended up being an inaccurate metric. Robodebt has led to serious financial and emotional consequences, with more than A$1.7 billion in debt collected from 430,000 people since its introduction in 2015.The 2020 Parliamentary Inquiry found that for a large number of respondents, it had systemically damaged their health and trust in government services. Ultimately, Robodebt was declared unlawful by the Federal Court of Australia and all money wrongly taken from individuals had to be repaid to those decision points with interest. The case illustrates a fundamental problem with the types of databases that digital technology can access, and how such AI-powered systems can cause distortions if they're built on dubious assumptions, especially when it comes to marginalised people.

In the UK, the Department for Work and Pensions has come under fire for its benefit fraud detection algorithm, which was found to unfairly target disabled people and certain

nationalities, leaving thousands without or with minimal benefits. The question of how these algorithms actually work has added a layer of distress for those affected, who had no choice but to know. It is yet another example of government being separated from the premise – the social protection function – where the solution (end-to-end) is more important than just getting the processes right.

At the same time, a project called "Meu INSS" (My Social Security), which is part of Brazil's goal to have all services digitized showed how heavy automation has its downfalls. But the project exacerbated a spike in automatic benefit denials: over 869,000 applications were denied by machine last year, more than double. These rising rejection rates - frequently for reasons valid only through inflexible, erroneous standards of eligibility, exacerbated by wider digital divisions in society - illustrate the potential injustice engendered by automation. According to the data from United Nations Development Programme, in Brazil 25% of inhabitants spend their life without access to internet.

Similar cases are emerging in these countries, underlining a trend of the state withdrawing from the provision of social protection. While automation helps in this regard, it often puts the burden on the most vulnerable to navigate difficult and opaque systems designed to force people to prove their eligibility for benefits. Changing hands for this responsibility can sometimes slow down the resolution of inappropriate denials, as systems prioritise process correctness over individual needs.

In addition, the consequences of corrupt databases cannot be overstated. Potentially damaging and unfair outcomes can result, as in the case of Australia and Brazil, with AI systems that may act on incorrect or insufficient data. Instead, social protection becomes an exact science dependent on AI - it should be a social science close to human needs.

These systems were operating long before the European regulation requiring a fundamental rights impact assessment was published. Such assessments could have prevented some of these in the future, had we only ensured that principles such as fairness, transparency and human dignity were applied to the assessment before AI systems found their way into social protection.

As shown in the introduction, the European regulation imposes a high-risk classification on the systems responsible for the implementation of rights to social security and social assistance benefits, mainly because they are not only systems for access to public assistance, but also support for people in situations of social risk: old age, pregnancy, illness, accidents at work, unemployment, homelessness. The aim of these systems is to implement legislative decisions (because the social risk to be protected is provided for by law) in very critical situations that affect the life, health and above all the livelihood of people in vulnerable situations.

However, without explicitly making this specific distinction, Annex III of the European regulation states more generally that these systems are high risk, but that it is necessary to monitor these experiences in order to review their operation.

As we can see, the European regulation on AI is indeed comprehensive and aims to strictly regulate AI systems related to eligibility for state support of social services and benefits, which have been praised and treated as high-risk AI systems. However, it is clear from the detailed requirements that the centrality lies in the subjection to the process, with other relevant issues

such as social and informational inequality, the systems' database and other intersectionalities that potentially affect the concreteness of the social protection that is expected for the ultimate purpose of safeguarding the dignity of the human person in a situation of social risk.

4. BIBLIOGRAPHY

AIAAIC. (2019). *Incident Number 189*. In: McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved from https://incidentdatabase.ai/cite/189 on July 24, 2024.

ALGORITHM WATCH. (2019). *Automating Society 2019* [online]. AlgorithmWatch. [Accessed: 24 July 2024]. Retrieved from https://algorithmwatch.org/en/automating-society-2019/ on July 24, 2024.

ATHERTON, D. (2021). *Incident Number 611*. In: Atherton, D. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved from https://incidentdatabase.ai/cite/611 on July 24, 2024.

BOOTH, R. (2023). UK warned over lack of transparency in use of AI to vet welfare claims. *The Guardian*. Retrieved from https://www.theguardian.com/politics/2023/sep/03/uk-warned-over-lack-transparency-use-ai-vet-welfare-claims

BRASIL. (2020). Decreto n.º 10.332, de 28 de abril de 2020. *Presidência da República*. Retrieved from https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/decreto/D10332.htm on July 7, 2024.

Défenseurs des droits. (2019). *Rapport: Dématérialisation et inégalités d'accès aux services publics*. Défenseurs des droits. Retrieved from https://www.defenseurdesdroits.fr/rapport-dematerialisation-et-inegalites-dacces-aux-services-publics-266 on July 24, 2024.

ERDBRINK, T. (2021). Government in Netherlands Resigns After Benefit Scandal. *The New York Times*. Retrieved from https://www.nytimes.com/2021/01/15/world/europe/dutch-government-resignation-rutte-netherlands.html

EUBANKS, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. ISBN: 978-1-250-07431-7.

European Commission. (2024). Digital Economy and Society Index. *European Commission*. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/desi-sweden on July 6, 2024.

IBGE. (2024). Pesquisa Nacional por Amostra de Domicílios: Acesso à Internet e à Televisão e Posse de Telefone Móvel Celular para Uso Pessoal. *IBGE*. Retrieved from https://biblioteca.ibge.gov.br/visualizacao/livros/liv101748.pdf on July 6, 2024.

INSS. (2024). *Portal do Meu INSS*. Retrieved from https://meu.inss.gov.br/#/login. on July 6, 2024.

INSS. (2024). *Benefícios Indeferidos*. Retrieved from https://www.gov.br/inss/on July 6, 2024.

La Quadrature du Net. (2022, October 19). *CAF: Le numérique au service de l'exclusion et du harcèlement des plus précaires*. Retrieved from https://www.laquadrature.net/2022/10/19/caf-le-numerique-au-service-de-lexclusion-et-du-harcelement-des-plus-precaires/ on July 7, 2024.

La Quadrature du Net (2024). Une numérisation du contrôle social discriminatoire. *Plein Droit, 2024/1 n° 140, 23-26*. Retrieved from https://droit.cairn.info/revue-plein-droit-2024-1-page-23?lang=fr on July 24, 2024.

MAXWELL, J. (2021). Judicial review and the Digital Welfare State in the United Kingdom and Australia. *English Law: Public Law (topic)*. https://doi.org/10.2139/ssrn.3896200.

McGREGOPR, S.; LAM, K. (2018). *Incident Number 101*. In: McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved from https://incidentdatabase.ai/cite/101 on July 24, 2024.

MISURACA, G., & VAN NOORDT, C. (2020). *Overview of the use and impact of AI in public services in the EU*. EUR 30255EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-19540-5, doi:10.2760/039619, JRC120399.

Offices of the Auditor General of Ontario. (2015). *Annual Report 2015*. Toronto (Ontario). Retrieved from https://www.auditor.on.ca/en/content/annualreports/arbyyear/ar2015.html on July 24, 2024.

PAUL, R. (2023). European artificial intelligence "trusted throughout the world": Risk-based regulation and the fashioning of a competitive common AI market. Regulation & Governance. Retrieved from https://doi.org/10.1111/rego.12563 on July 24, 2024.

PNUD. (2024). Relatório de Desenvolvimento Humano 2023-2024. *Programa das Nações Unidas para o Desenvolvimento*. Retrieved from https://hdr.undp.org/system/files/documents/global-report-document/hdr2023-24overviewen.pdf on July 6, 2024.

ROBERTS, A. (2023). Michigan court approves $20M settlement in 2015 unemployment lawsuit. Detroit Free Press. https://www.freep.com/story/money/business/2023/01/24/michigan-court-approves-20m-settlement-2015-unemployment-lawsuit/69834046007/

TWEEDE KAMER DER STATEN-GENERAAL. (2020). *Eindverslag Parlementaire Ondervragingscommissie Kinderopvangtoeslag*. Retrieved from https://www.tweedekamer.nl/sites/default/files/atoms/files/20201217_eindverslag_parlementaire_ondervragingscommissie_kinderopvangtoeslag.pdf on July 24, 2024.

Services Australia. (2024). *Centrelink*. Retrieved from https://www.servicesaustralia.gov.au/. on July 6, 2024.

Capítulo 20

LA TRANSFORMACIÓN DIGITAL DE LOS DERECHOS DE SEGURIDAD SOCIAL EN BRASIL ANTE LA INTELIGENCIA ARTIFICIAL Y LOS RIESGOS PARA LA PROTECCIÓN SOCIAL

Mariah Brochado

Chair of Philosophy of Technology and Digital Rights (UFMG). Visiting Professor at the Leibniz-Institut für Medienforschung | Hans-Bredow-Institut. President of the Artificial Intelligence in Law of Minas Gerais' Bar commission.

Lucas Porto

PhD in Law (UFMG). Senior researcher at the Chair of Philosophy of Technology and Digital Rights (Philotech – UFMG). Socialenvironmental Scientist. Lawyer.

Roberto de Carvalho Santos

PhD candidate at Law (UFMG). President of Brazil's Social Security Studies Institute (IEPREV).

RESUMEN: La digitalización de los derechos sociales en Brasil, particularmente en el ámbito de la seguridad social, ha avanzado significativamente con la implementación del programa «INSS Digital» en 2017 y del portal «Meu INSS». Estos desarrollos han reducido la dependencia de los servicios presenciales y han incrementado el uso del análisis automatizado de las prestaciones, conforme a lo establecido por la Ley n.º 13.846 de 2019. Aunque la digitalización ha mejorado la capacidad para procesar solicitudes, también ha suscitado inquietudes sobre la privacidad, la seguridad de los datos y la eficiencia del servicio, además de haber incrementado el número de solicitudes rechazadas. Para 2023, muchos procesos se completaban de manera automática, lo que intensificó las críticas por la falta de transparencia y la percepción de injusticia. La automatización enfrenta desafíos relacionados con la equidad y el acceso, particularmente para los sectores más vulnerables. Los debates políticos han girado en torno a la protección de datos y la no discriminación, abordando preocupaciones sobre el sesgo algorítmico y la transparencia. Garantizar los derechos sociales de manera justa y transparente es fundamental para el desarrollo nacional y para corregir prácticas de mercado que resulten perjudiciales.

PALABRAS CLAVE: Digitalización; Seguridad social; Automatización; Transparencia; Derechos sociales

1. INTRODUCCIÓN

En los albores del siglo XXI, en un contexto histórico en el que el Estado ha adoptado una nueva fisiología y nuevas facetas, particularmente en lo que respecta a su papel en la garantía de los derechos sociales, los objetos técnicos también se han reformulado, configurando lo que hoy se denomina gobierno digital. Este se basa en herramientas tecnológicas para gestionar las actividades emergentes del Estado, orientadas principalmente a la prestación de servicios.

Es en este contexto donde se retoma el uso de sistemas de inteligencia artificial (IA) en la gestión de políticas públicas, particularmente aquellas orientadas a la provisión de derechos sociales. En especial, el derecho a la seguridad social, una categoría de derecho social que conforma el trípode de la seguridad social, está marcado por el impasse de la legibilidad (Das; Poole, 2004, p. 9), el cual atraviesa al individuo. Esta situación dificulta el acceso a los derechos por parte de individuos en situación de riesgo social. En otras palabras, el individuo se ve constreñido por la producción de lenguajes y formas de conocimiento propias del Estado, que se emplean como instrumentos de clasificación y regulación de las colectividades. Estos mecanismos, a menudo burocráticos y lentos, afectan especialmente a aquellos individuos que, en general, presentan un bajo nivel educativo y, en ese momento, se encuentran aún más vulnerables debido a su situación de incapacidad laboral o avanzada edad. Este escenario se agrava con la actualización de las prácticas de acceso a los derechos de seguridad social en Brasil, impulsada por la promesa de un procesamiento ágil de la información, particularmente a través de lo que hoy se conoce como Big Data: sistemas que almacenan y procesan grandes volúmenes de datos extremadamente diversos, generados, capturados y procesados a gran velocidad (Chen; Chiang; Storey, 2012, p. 1165). Este es el contexto de este estudio, cuyo objetivo es introducir a los lectores en el proceso de digitalización de los derechos de seguridad social en Brasil.

2. NOCIONES INTRODUCTORIAS SOBRE LA INTELIGENCIA ARTIFICIAL

La inteligencia artificial (IA) es un campo multifacético que, según revela la literatura actual, puede ser explorado desde cuatro ángulos distintos: (i) como área de la informática que se entrelaza con la psicología y la neurociencia para influir en los procesos computacionales; (ii) como una manifestación de la algorítmica, que emplea métodos matemáticos para modelar proyectos computacionales; (iii) como evolución de la cibernética, que integra la noción de redes y conexiones para simular funciones cerebrales y ampliar la capacidad computacional (Brochado, 2023, p. 273-274); y (iv) como un campo práctico, enfocado en mejorar las capacidades computacionales mediante software que optimiza el hardware disponible (Brochado, 2023, p. 287-298).

Es en la conceptualización de la inteligencia artificial (IA) desde una perspectiva pragmática donde han emergido las diversas escuelas de aprendizaje automático. Según Domingos

(2015, p. xvii), varias corrientes de pensamiento definen el proceso de aprendizaje en IA. Los principales grupos en este campo son los simbolistas, los conexionistas, los evolucionistas, los bayesianos y los analogistas. Cada uno de estos grupos se adhiere a un conjunto fundamental de creencias, se enfoca en un problema específico al que otorgan prioridad, y ha desarrollado una solución particular para dicho problema, basándose en conceptos derivados de sus respectivas disciplinas científicas aliadas. De este modo, cada escuela ha creado un "algoritmo maestro" (Domingos) que encapsula su enfoque distintivo.

La aparición de sesgos potencialmente indeseables, junto con el discurso que celebra la sustitución de la toma de decisiones humanas —caracterizada por una capacidad de análisis matizada— por meras variables gestionadas por algoritmos de aprendizaje automático, plantea un nuevo desafío al marco jurídico. En esta línea, Hildebrandt (2016) sostiene que el Derecho ofrece orientación sobre las consecuencias jurídicas de nuestras acciones, indicándonos qué podemos esperar, como, por ejemplo, las obligaciones exigibles relacionadas con el pago de indemnizaciones o los derechos de transferencia de propiedad sobre determinados bienes. Al hacerlo, el Derecho configura las dimensiones temporal y espacial del mundo en el que vivimos y del cual dependemos. Por tanto, estudiar el Derecho no se limita a la mera recuperación de información; implica comprender la naturaleza única de las fuentes jurídicas, las cuales están intrínsecamente vinculadas a la autoridad coercitiva. Sin embargo, la delegación de decisiones operadas por seres humanos a entidades maquínicas sigue careciendo de un reconocimiento exhaustivo, en el ámbito formativo, del carácter performativo inherente a la práctica jurídica, el cual define la estructura de las decisiones emitidas.

3. LA PROTECCIÓN DE LOS DERECHOS SOCIALES EN LA REGULACIÓN DE LA INTELIGENCIA ARTIFICIAL

Aunque se reconoce ampliamente que el rápido desarrollo e implantación de las tecnologías centradas en los datos tiene efectos transformadores, los detalles de estos efectos siguen siendo objeto de debate. Las preocupaciones iniciales se centraron en la amplia recopilación de datos, haciendo hincapié en cuestiones de vigilancia y privacidad, especialmente en el ámbito de lo que se conoce como capitalismo de vigilancia (Brochado, 2023, pp. 299-301). Estas discusiones han puesto de manifiesto las deficiencias de las leyes existentes y han contribuido a un debate en curso sobre la necesidad de mejorar la protección de la privacidad y los datos personales, así como de mejorar la supervisión de la gestión de datos por parte de empresas y entidades gubernamentales. Muchas de estas cuestiones se han abordado en la normativa de protección de datos, cuyo objetivo es reforzar la protección de los derechos fundamentales en medio de la rápida evolución de las tecnologías y los servicios digitales.

La preocupación por la privacidad ha sido predominante en relación con las tecnologías de optimización; sin embargo, actualmente existen programas de investigación que se centran en cuestiones como la clasificación automatizada y los sesgos presentes en los datos y algoritmos, los cuales pueden derivar en discriminación (Lee, 2019). La privacidad y la no dis-

criminación han emergido como conceptos organizativos clave en los debates políticos sobre estas tecnologías. No obstante, al evaluar el potencial transformador de dichas tecnologías, las políticas de privacidad y no discriminación también enfrentan limitaciones. La implementación de estas prioridades ha sido objeto de críticas, ya que ha dado lugar a soluciones de diseño orientadas a resolver problemas a través de iniciativas como la «privacidad por diseño» o la mitigación de sesgos. Si bien estas iniciativas pueden ofrecer ciertos beneficios, rara vez cuestionan la naturaleza misma del sistema de IA o sus lógicas operativas, lo que limita su capacidad para generar soluciones duraderas. Como aclara Hoffmann:

Debemos resistirnos activamente a la idea de que los datos y los algoritmos se limitan a informar, apoyar o emitir decisiones que afectan la distribución de bienes específicos. En particular, es crucial enfrentarnos directamente al papel que desempeñan los datos y los algoritmos en la mediación activa y en la normalización de discursos y condiciones sociales que, en primera instancia, condicionan las decisiones sobre dichas distribuciones. Kate Crawford y Vladan Joler (2018) plantean esta cuestión de manera más drástica: «Muchos de los supuestos sobre la vida humana que integran los sistemas de aprendizaje automático son estrechos, normativos y están plagados de errores. Sin embargo, estos supuestos se están inscribiendo y construyendo en un nuevo mundo, y desempeñarán un papel cada vez más importante en la forma en que se distribuyen las oportunidades, la riqueza y el conocimiento». (Hoffmann, 2019).

Además, al centrarse exclusivamente en los derechos individuales, se dejan de reconocer las transformaciones estructurales introducidas por las tecnologías de optimización y su impacto en los derechos sociales, como el papel del trabajo y la protección del Estado de bienestar. Esto ocurre a pesar de que dichas transformaciones no forman parte de los enfoques dominantes relacionados con los datos y las infraestructuras informáticas. El debate sobre la precariedad en la intersección entre las tecnologías de optimización y el trabajo también amplía las discusiones sobre el futuro del Estado de bienestar. Esta cuestión no solo abarca la seguridad de los derechos laborales o las garantías de ingresos, sino que cada vez más examina cómo las infraestructuras de datos influyen en los servicios públicos, incluidos los controles de elegibilidad, las evaluaciones de riesgos y la elaboración de perfiles (Ewbanks, 2018, p. 54).

En un informe a la Asamblea General sobre la pobreza extrema y los derechos humanos, estas tendencias fueron caracterizadas como la aparición del Estado de bienestar digital, una realidad que está emergiendo en varios países alrededor del mundo. Según Van Zoonen (2020), esta transición hacia el uso de datos en las políticas sociales no está exenta de riesgos. Virginia Eubanks, por ejemplo, documenta numerosos casos de automatización y datificación negligente en las políticas sociales de varios estados de EE. UU., los cuales han dejado a millones de personas injustamente acusadas de fraude y privadas de sus prestaciones. Eubanks concluye, tras años de entrevistas y observación exhaustiva, que las tecnologías de datos y los algoritmos han creado una "pobreza digital," en la que los grupos ya desfavorecidos están sometidos a un mayor control y vigilancia que nunca. Eubanks (2018, p. 13) afirma:

Al igual que con las anteriores innovaciones tecnológicas en la gestión de la pobreza, el seguimiento digital y la toma de decisiones automatizada ocultan la pobreza del público

profesional de clase media, proporcionando a la nación la distancia ética necesaria para tomar decisiones deshumanizadoras: quién recibe alimentos y quién pasa hambre, quién tiene vivienda y quién queda sin hogar, y qué familias son disgregadas por el Estado. La "pobreza digital" se inserta en una larga tradición estadounidense, en la que gestionamos a los pobres de manera individual, eludiendo así nuestra responsabilidad compartida de erradicar la pobreza. (Ewbanks, 2018, p. 13).

Uno de los principales argumentos del Estado para adoptar sistemas de inteligencia artificial (IA) en la provisión de derechos sociales es la lucha contra el fraude y los errores, los cuales pueden representar grandes sumas de dinero. En consecuencia, el Informe de la ONU revela que muchos de los sistemas digitales de asistencia social implementados han sido específicamente diseñados para mejorar la capacidad de cruzar datos procedentes de diversas fuentes con el fin de detectar fraudes e irregularidades por parte de los solicitantes de ayuda social. Sin embargo, los estudios de campo realizados por el Relator Especial de la ONU, junto con otros casos analizados, indican que a menudo se exagera la magnitud de estos problemas. En ocasiones, se presta una atención desproporcionada a este aspecto del sistema de bienestar (Naciones Unidas, 2019, p. 10).

El cálculo del riesgo también ocupa un lugar central en los objetivos de los sistemas de bienestar, y las tecnologías han alcanzado un alto nivel de sofisticación en este ámbito. Además de la detección y prevención del fraude, la protección del trabajo infantil ha sido un punto clave, como lo ilustran ejemplos en países como Dinamarca, Nueva Zelanda, el Reino Unido y Estados Unidos. Estas tecnologías también se emplean para determinar la elegibilidad y calcular la cuantía de las prestaciones por desempleo. Por ejemplo, mientras que en Polonia dicho sistema ha sido declarado inconstitucional, en Austria un sistema basado en algoritmos continúa evaluando a los solicitantes desempleados para decidir la cuantía de la ayuda que recibirán del Estado (Naciones Unidas, 2019, p. 11).

Muchos otros ámbitos del Estado de bienestar también se ven afectados por tecnologías que evalúan riesgos y categorizan necesidades. Aunque estos métodos ofrecen numerosas ventajas, especialmente en términos de rapidez y ahorro de costes, es fundamental tener en cuenta los problemas que pueden surgir. En primer lugar, basar los derechos de un individuo en predicciones derivadas del comportamiento de un grupo poblacional plantea una serie de dilemas éticos y prácticos. En segundo lugar, el funcionamiento de estas tecnologías y los métodos utilizados para generar puntuaciones o clasificaciones suelen ser opacos, lo que complica la rendición de cuentas por parte de gobiernos y entidades privadas ante posibles violaciones de derechos. Por último, las prácticas de puntuación de riesgos y categorización de necesidades pueden perpetuar o incluso agravar las desigualdades y la discriminación preexistentes (Naciones Unidas, 2019, p. 10).

Estas diversas preocupaciones subrayan la importancia de los derechos sociales en el contexto de la digitalización y la introducción de tecnologías de optimización, aunque frecuentemente no se abordan de manera directa. Si bien la privacidad y la protección de datos en los ámbitos del trabajo y el bienestar social forman parte de estos debates, los derechos sociales rara vez han sido el foco central. Todavía no se sabe con certeza cómo integrar eficazmente

estas cuestiones en los debates políticos ni cómo influir en las agendas legislativas en relación con las infraestructuras de datos y las tecnologías emergentes. En consecuencia, el proceso político relacionado con la tecnología tiende a centrarse en la regulación de los riesgos y en la asignación de recursos para la innovación, priorizando cuestiones procedimentales y presupuestarias por encima de consideraciones sobre la naturaleza del trabajo o la sostenibilidad de los servicios públicos. Este enfoque favorece a ciertos actores y discursos tecnocéntricos que suelen dominar las consultas políticas y moldear la agenda de los debates.

4. DERECHO DE LA SEGURIDAD SOCIAL E INTELIGENCIA ARTIFICIAL: LA TRANSFORMACIÓN DIGITAL DEL INSS Y LOS RIESGOS PARA LA PROTECCIÓN SOCIAL

En Brasil, el Instituto Nacional de Seguridad Social (INSS), la autoridad federal responsable de gestionar el reconocimiento de los derechos de los asegurados del Régimen General de la Seguridad Social (RGPS), es actualmente el mayor organismo público de distribución de ingresos en América Latina. En 2017, se lanzó la estrategia «INSS Digital», formalizada por la Instrucción Normativa 96/PRES/INSS (Brasil, 2018), la cual estableció el portal «Meu INSS» como el principal medio para la emisión de declaraciones y la solicitud de servicios ante el Instituto, relegando la atención presencial a un papel secundario, limitado a citas programadas previamente. Posteriormente, el Poder Ejecutivo avanzó en esta dirección mediante la Medida Provisional n.º 871 de 2019, que posteriormente fue convertida en la Ley n.º 13 846 del 18 de junio de 2019, ampliando la digitalización de los servicios del INSS. Esta ley también introdujo una bonificación para el análisis de prestaciones con posibles irregularidades: una para los funcionarios que revisan procesos fuera del horario laboral, y otra para los peritos médicos en casos de prestaciones por invalidez.

En las más de 1.500 oficinas del INSS distribuidas por todo el país —muchas de ellas de reciente construcción en el marco del Plan de Ampliación y Reforma de la Estructura de Servicios del INSS, abandonado en 2016—, la atención presencial al asegurado/usuario se limita a unos pocos servicios específicos, previa reserva. Entre estos servicios se encuentran la peritación médica, la valoración social y el cumplimiento de ciertos requisitos administrativos. No obstante, incluso estos servicios presenciales limitados se ven amenazados por la actual estrategia de gestión. Cabe destacar también la Ordenanza DIRBEN/INSS n.º 978, del 4 de febrero de 2021, que permite la evaluación social de personas con discapacidad mediante videoconferencia en todo el país, siempre que así lo desee el beneficiario. En cuanto a los exámenes médicos, la Ordenanza Conjunta DIRBEN/INSS/SPMF/SPREV/MTP n.º 1, del 26 de enero de 2022, delineó el proceso para la realización de Exámenes Médicos por Teleevaluación (PMUT) como parte de un proyecto piloto.

Cabe mencionar también la violación del sigilo personal de los asegurados hiposuficientes y vulnerables introducida por el artículo 124-B de la Ley n.º 8.213/91, mediante la Ley n.º 13.846/19. Esta disposición, incluida en la legislación conocida como «operación peine»,

no pretende beneficiar a los asegurados. Al contrario, pretende cruzar datos para identificar, a toda costa, indicios de irregularidades en un contexto en el que el Estado presume la mala fe de los ciudadanos.

La responsabilidad procedimental administrativa ha recaído íntegramente sobre el solicitante, quien, al solicitar un servicio, no anticipa que al seleccionar la opción «Deseo seguir el progreso de mi solicitud a través de Mi INSS o del teléfono 135», estará eximiendo al INSS de la obligación de notificarle por correo, el medio tradicionalmente utilizado, sobre la programación de los reconocimientos médicos o la aprobación de su solicitud.

Durante la prueba de concepto y la fase piloto del nuevo modelo de servicio del INSS, se recopilaron sugerencias de los funcionarios con vistas a mejorar los sistemas, la legislación y los procedimientos. Para evaluar la satisfacción de estos empleados, se realizaron dos encuestas y se crearon canales de comunicación, como grupos de WhatsApp y el correo electrónico inssdigital@inss.gov.br, para facilitar el intercambio de opiniones. Además, en la fase PDCA (Planificar, Hacer, Comprobar, Actuar) del experimento, con vistas a una expansión nacional segura, se publicaron los datos relevantes en la página intraprev de la intranet del INSS, que incluye los principales hitos del proyecto, reglamentos ajustados, guías prácticas para apoyar las actividades de los funcionarios y testimonios de los participantes, reforzando los mecanismos de transparencia y control social (Brasil, 2017, p. 232-233).

En mayo de 2023, el INSS registró un número récord de decisiones automáticas sobre solicitudes, el más alto desde que se introdujo la IA para analizar las prestaciones. En ese mes, el 42 % de los procesos se resolvieron de manera automática, lo que representó más de 222.000 prestaciones (Brasil, 2023). Por otro lado, un boletín divulgado en mayo de 2022 informaba de una larga espera para que las solicitudes fueran analizadas, con un promedio de 115 días de espera, y revelaba que más de 1,14 millones de solicitudes de beneficios fueron denegadas en todo el país durante el primer trimestre de ese año (Cresce..., 2022). Además, entre 2012 y 2018, el instituto rechazó una media de 3,4 millones de prestaciones anuales en Brasil. Desde 2019, esa cifra ha aumentado a una media de 4,4 millones de prestaciones denegadas por año. Según el último Boletín Estadístico de la Seguridad Social (BEPS), publicado en junio de 2023, hubo un aumento de alrededor del 10 % en las denegaciones de beneficios en general, y de beneficios por discapacidad en particular, tanto en comparación con mayo de 2023 como con el mismo periodo del año anterior (Brasil, 2023).

Esta situación se ha deteriorado aún más tras la implementación de un proceso de automatización en el análisis previo para la concesión de prestaciones de la Seguridad Social, iniciado en mayo de 2022, después de la pandemia de COVID-19. Desde entonces, se ha observado un incremento del 5 % en las denegaciones automáticas de prestaciones, y se estima que la automatización de los análisis alcanzará el 50 % en 2026 (Gercina, 2023). Además, la implementación del análisis automático mediante IA entre 2022 y 2023 elevó la tasa de recursos denegados del 17 % al 36 % (Alvarenga, 2023).

En junio de 2024, la Procuraduría General de la Unión (AGU) tiene previsto lanzar «Pacífica», un proyecto piloto que emplea inteligencia artificial (IA) para revisar las solicitudes de prestaciones denegadas por el Instituto Nacional de Seguridad Social (INSS). El objetivo de

esta revisión automatizada es verificar la validez de las solicitudes y reducir la necesidad de que los solicitantes recurran al sistema judicial, aliviando así su carga. Tras la denegación de una solicitud por parte del INSS, la IA analizará el caso para determinar la viabilidad de la solicitud, fomentando una interacción preliminar entre el solicitante y la AGU antes de que se recurra a los tribunales. La AGU utilizará los datos de las solicitudes y la IA, que opera de acuerdo con normas específicas de toma de decisiones, para evaluar si la solicitud puede ser aceptada y, en consecuencia, concedida la prestación. No obstante, existen preocupaciones dentro de la autarquía de que la IA pueda reproducir injusticias debido a las normas implementadas. Por este motivo, agentes de la AGU también participarán en el proceso, garantizando una evaluación más justa y equilibrada (Portal Contábeis, 2024).

Esta medida responde a la intención de corregir las deficiencias identificadas por una auditoría de la Oficina del Interventor General. La auditoría evaluó los procesos entre 2021 y 2023, aunque la automatización comenzó en 2017. Desde entonces, el número de prestaciones analizadas automáticamente ha aumentado considerablemente, y para 2022 ya se contaban más de 1,3 millones de análisis de este tipo. De estos, 869.000 resultaron en la denegación de la solicitud, es decir, aproximadamente dos de cada tres. Esta proporción es significativamente superior a la observada en los análisis manuales, en los que se rechazó el 50 % de las solicitudes. Se ha detectado que la tasa de rechazo ha aumentado en paralelo al incremento del volumen de análisis. Por ejemplo, en 2021, cuando se evaluaron automáticamente 490.000 solicitudes, solo se rechazó el 41 % (G1, 2023).

Con el creciente uso de sistemas de IA para tomar decisiones técnicas en la Administración Pública, destaca el análisis de Brochado (2023, pp. 528-529) sobre la urgente necesidad de debatir programas de aprendizaje automático. Según el autor, el sistema probabilístico que permite la autonomía operativa de estas máquinas es incontrolable y es en esta naturaleza de los programas donde radica su eficacia distintiva. Brochado (2023, pp. 502/512-513) también señala la auditabilidad de estos sistemas como uno de los mayores desafíos para la legislación brasileña, y destaca la importancia de comprender claramente el proceso de adopción de las normas decisorias y de mantener la transparencia en las técnicas estadísticas utilizadas para entrenar los algoritmos.

El lenguaje, que puede ser traducido y procesado por máquinas, transmite patrones de información organizados en «paquetes» que pueden ser utilizados por diversos agentes de control social, alineados con los dispositivos burocráticos del Estado. En paralelo, Frazão (2021) analiza el «derecho a la explicación» en Brasil, basado en el artículo 20 de la Ley General de Protección de Datos (LGPD), que otorga a los ciudadanos el derecho a solicitar una revisión de las decisiones basadas exclusivamente en el tratamiento automatizado de datos personales que afecten sus intereses, incluidas aquellas que definen perfiles personales y profesionales. De manera similar, el artículo 5 de la Carta Ética Europea sobre el Uso de la Inteligencia Artificial en los Sistemas Judiciales establece que los profesionales del sistema judicial deben tener la capacidad de revisar las decisiones y los datos utilizados para emitir un veredicto, sin estar necesariamente vinculados a ellos, tomando en cuenta las particularidades de cada caso (Comisión Europea para la Eficiencia de la Justicia, 2018).

A partir del 10 de abril de 2024, los ciudadanos que utilicen un smartphone para contactar con la Central 135 y solicitar servicios de la Seguridad Social tendrán acceso a un menú de autoservicio digital en la pantalla de su dispositivo móvil, previo a la realización de la llamada (INSS, 2024). Este menú ha sido diseñado basándose en los temas más frecuentemente solicitados en la Central. La implementación de esta funcionalidad se llevará a cabo de manera gradual, comenzando con una fase de prueba. No obstante, cabe destacar que, una vez más, se trata de una iniciativa que no ha sido sometida a pruebas con el público objetivo al que está destinada.

En contraposición a esta perspectiva, la realidad de las decisiones automatizadas en el INSS parece manifestar diferencias significativas. Conforme a una nota técnica divulgada en junio de 2023, el INSS afirma que, con el propósito de garantizar la calidad del servicio y optimizar la experiencia del usuario, su equipo tecnológico lleva a cabo una supervisión técnica continua de los procesos para perfeccionar los sistemas. Esta declaración sugiere que el control de los datos está centralizado en la Administración, la cual asume la responsabilidad de su mejora. Adicionalmente, la Dirección de Tecnología e Innovación (DTI) del organismo puntualiza que las plataformas de automatización incorporan las normativas de seguridad social establecidas por ley y los sistemas del INSS en un mecanismo que promueve la celeridad en el análisis y la resolución de las solicitudes, considerando también las observaciones formuladas por los ciudadanos al solicitar el beneficio (Brasil, 2023).

5. CONCLUSIONES

La transformación digital de los derechos sociales en Brasil, con especial énfasis en el ámbito de la seguridad social, ha estado caracterizada por la implementación de la estrategia denominada «INSS Digital», cuyo inicio se remonta a 2017. Esta iniciativa introdujo el portal «Meu INSS» como principal plataforma para la solicitud de servicios y la emisión de declaraciones, lo que ha conllevado una reducción significativa en la dependencia de la atención presencial, la cual ahora requiere de una reserva previa. El proceso de digitalización experimentó un avance notable con la promulgación de la Ley n.º 13.846 de 2019, la cual no solo intensificó el análisis automatizado de las prestaciones, sino que también incorporó incentivos destinados a que tanto funcionarios como peritos médicos identificasen posibles irregularidades en el sistema.

La transformación digital no solo ha ampliado la capacidad de tramitación de solicitudes del INSS, sino que también ha suscitado preocupaciones significativas en materia de privacidad, seguridad de datos y eficiencia del servicio. La automatización ha intensificado el rechazo de solicitudes, evidenciándose un incremento en las denegaciones de prestaciones. En 2023, una proporción considerable de los procesos se completaron de manera automática, lo que ha exacerbado las críticas sobre la falta de transparencia y el riesgo de inequidad en la tramitación de las solicitudes.

La automatización de la Seguridad Social en Brasil se enfrenta a desafíos en términos de equidad y acceso a los derechos sociales. El avance de la digitalización y la automatización

ha puesto de manifiesto la necesidad imperiosa de debatir y reevaluar la interacción entre la tecnología y los derechos sociales, particularmente para garantizar que la implementación de nuevas tecnologías no excluya ni perjudique a los segmentos más vulnerables de la población. La eficacia de estas transformaciones continúa dependiendo de un compromiso sostenido para perfeccionar los sistemas y asegurar que todos los ciudadanos puedan acceder a sus derechos de manera justa y transparente.

El enfoque de los debates políticos sobre tecnología se ha centrado en la protección de datos y la no discriminación, áreas cruciales de intervención debido a las preocupaciones sobre el procesamiento de datos, el sesgo algorítmico y la transparencia de los modelos computacionales. El debate sobre la IA frecuentemente gravita en torno a la discriminación y el sesgo algorítmico, lo que ha suscitado críticas sobre la superficialidad de los debates en materia de inclusión y diversidad, así como sobre las respuestas excesivamente centradas en la tecnología frente a las desigualdades. Estas cuestiones están condicionadas por factores como la implicación de las empresas y las narrativas de los medios de comunicación.

A pesar de los retos, los derechos sociales continúan siendo fundamentales para el desarrollo de cualquier nación, esenciales para corregir las prácticas de mercado perjudiciales y para configurar los marcos reguladores. Se considera prioritario abordar cuestiones estructurales como las instituciones, los mecanismos de redistribución y el control de las infraestructuras y las finanzas públicas, en conjunción con la regulación de las propuestas relativas a los derechos sociales.

6. BIBLIOGRAFÍA

ALVARENGA, Laura. INSS: Brasileiros estão DESESPERADOS com aumento de indeferimentos de benefícios; confira o motivo. FDR, 1 ago. 2023. Disponível em: https://fdr.com.br/2023/08/01/inss-brasileiros-estao-desesperados-com-aumento-de-indeferimentos-de-beneficios-confira-o-motivo/. Acesso em: 1 abr. 2024.

BRASIL. INSS Digital: uma nova forma de atender, 2017. https://repositorio.enap.gov.br/bitstream/1/4144/1/INSS%20Digital_Uma%20nova%20forma%20de%20atender.pdf.

BRASIL. Instituto Nacional do Seguro Social (INSS*)*. Diretoria de Tecnologia da Informação – DTI. Automação é aliada na agilização das decisões do INSS, 19 jun. 2023. Disponível em: https://www.gov.br/inss/pt-br/assuntos/noticias/automacao-e-aliada-na-agilizacao-das-decisoes-do-inss. Acesso em 7 abr. 2024.

BRASIL. Instituto Nacional do Seguro Social (INSS). Instrução Normativa nº 96/PRES/INSS. 2018.

BRASIL. Instituto Nacional do Seguro Social (INSS). Portaria DIRBEN/INSS Nº 978, de 4 de fevereiro de 2021. Diário Oficial da União, Brasília, DF, 5 fev. 2021.

BRASIL. Instituto Nacional do Seguro Social (INSS). Boletim Estatístico da Previdência Social – BEPS, v. 28, n. 6, jun., 2023. Disponível em: https://www.gov.br/previdencia/pt-br/assuntos/previdencia-social/arquivos/beps062023_final-1.pdf. Acesso em: 8 abr. 2024.

BRASIL. Instituto Nacional do Seguro Social (INSS). Portaria Conjunta DIRBEN/INSS/SPMF/SPREV/MTP Nº 1, de 26 de janeiro de 2022. Diário Oficial da União, Brasília, DF, 27 jan. 2022. Seção 1

BRASIL. Lei nº 13.846, de 18 de junho de 2019. Diário Oficial da União, Brasília, DF, 19 jun. 2019. Seção 1, p. 1.

BRASIL. Medida Provisória nº 871, de 18 de janeiro de 2019. Diário Oficial da União, Brasília, DF, 18 jan. 2019. Seção 1, p. 1.

BROCHADO, Mariah. Inteligência Artificial no Horizonte da Filosofia da Tecnologia: técnica, ética e direito na era cybernética. São Paulo: Editora Dialética, 2023.

CHEN; CHIANG; STOREY. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, v. 36, n. 4, p. 1165, 2012.

COMISIÓN EUROPEA PARA LA EFICIENCIA DE LA JUSTICIA. Carta ética europeia sobre o uso da inteligência artificial em sistemas judiciais e seu ambiente. Tradução de Teresa Germana Lopes de Azevedo. ln: ENCONTRO NACIONAL DE JUÍZES ESTADUAIS, 7,, 2019, Foz do Iguaçu. Foz do Iguaçu: ENAJE, 2019.

COMISIÓN EUROPEA. White Paper on Artificial Intelligence—A European approach to excellence and trust (White Paper COM(2020) 65 final). Comissão Europeia. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

CRESCE número de indeferimentos de benefícios no INSS; saiba o que fazer após receber negativa do órgão. O Globo, 21 jul. 2022. Disponível em: https://extra.globo.com/economia-e-financas/cresce-numero-de-indeferimentos-de-beneficios-no-inss-saiba-que-fazer-apos-receber-negativa-do-orgao-25516368.html. Acesso em: 1 abr. 2024.

DAS, Veena; POOLE, Deborah. Anthropology in the margins of state, 2004

DOMINGOS, Pedro. The master algorithm: how the quest for the ultimate learning machine will remake our world. Nova Iorque: Perseus Books, 2015.

EWBANKS, Virginia. Automating inequality: how high-tech tools profile, police, and punish the poor. Nova Iorque: St. Martin's Press, 2018.

FRAZÃO, Ana. Decisões algorítmicas e direito à explicação. Jota, 24 nov. 2021. Disponível em: https://www.jota.info/opiniao-e-analise/colunas/constituicao-empresa-e-mercado/decisoes-algoritmicas-e-direito-a-explicacao-24112021

G1. CGU: Análise automática de benefícios do INSS tem mais negativas e risco de decisões indevidas. Disponível em: https://g1.globo.com/economia/noticia/2023/11/09/cgu-analise-automatica-de-beneficios-do-inss-tem-mais-negativas-e-risco-de-decisoes-indevidas.ghtml. Acesso em: 10 out. 2023.

GERCINA, Cristiane. INSS: robôs negam aposentadoria em seis minutos. Estado de Minas, 31 jul. 2023. Disponível em: https://www.em.com.br/app/noticia/economia/2023/07/31/internas_economia,1538168/inss-robos-negam-aposentadoria-em-seis-minutos.shtml. Acesso em: 7 abr. 2024.

HILDEBRANDT, Mireille. Law as Information in the Era of Data-Driven Agency. The Modern Law Review, v. 79, n. 1, p. 1–30, jan. 2016.

HOFFMANN, A. L. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society, v. 22, n. 7, p. 900–915, 7 jun. 2019.

INSTITUTO NACIONAL DO SEGURO SOCIAL (INSS). Ligação para a Central 135 tem menu digital na tela do celular a partir de amanhã. Disponível em: https://www.gov.br/inss/pt-br/assuntos/ligacao-para-a-central-135-tem-menu-digital-na-tela-do-celular-a-partir-de-amanha.

LEE, Kai-Fu. Inteligência artificial: como os robôs estão mudando o mundo, a forma como amamos, nos comunicamos e vivemos. Tradução de Marcelo Barbão. Rio de Janeiro: Globo Livros, 2019.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. Report of the Special Rapporteur on extreme poverty and human rights (A/74/493). UN Special Rapporteur on extreme poverty and human rights.

PORTAL CONTÁBEIS. IA revisará benefícios negados oferecidos pelo INSS. Disponível em: https://www.contabeis.com.br/noticias/64606/ia-revisara-beneficios-negados-oferecidos-pelo-inss/. Acesso em: 10 abr. 2024.

UNIÓN EUROPEA. Decisão (UE) 2022/2481 do Parlamento Europeu e do Conselho de 14 de dezembro de 2022 que estabelece o programa Década Digital para 2030. https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A32022D2481

VAN ZOONEN, L. Data governance and citizen participation in the digital welfare state. Data & Policy, v. 2, p. e10, 2020.

ISBN: