Deusto Journal of Human Rights

Revista Deusto de Derechos Humanos

http://djhr.revistas.deusto.es/
DOI: https://doi.org/10.18543/djhr

ISSN 2530-4275
ISSN-e 2603-6002

No. 14 Year / Año 2024

DOI: https://doi.org/10.18543/djhr142024

Inteligencia Artificial y Derechos Humanos: desafíos y oportunidades en la era digital

Contents / Índice

Introducción al monográfico, *José Miguel Iturmendi Rubia*

I. ARTICLES / ARTÍCULOS

Human rights, vulnerability and artificial intelligence: an analysis in constitutional perspective
Jorge Castellanos Claramunt

AI in supply chains: freedom from slavery revisited
Migle Laukyte & Lorena María Arismendy Mengual

The systematics of the European Artificial Intelligence Act in the context of the fundamental rights of the Union: the myth of the digital constitutionalism
Ainhoa Lasa López

Facing fundamental rights in the age of preventive *ex ante* AI: a contemporary form of discrimination
Mª Teresa García-Berrio Hernández

Opportunities and challenges of AI chatbots for digital youth information, advice, and counselling services in Europe
Alonso Escamilla & Paula Gonzalo

The human right to participate and its connection to artificial intelligence
María Dolores Montero Caro

Towards a better protection of human rights through the use of AI and related technologies in budgeting and auditing of public expenditure
María Amparo Grau Ruiz

Ética, desafíos y riesgos del acceso a la justicia algorítmica
José Carlos Fernández Rozas

Ética en crisis: El impacto de la carrera armamentística de las armas autónomas en nuestros valores morales
Jorge Couceiro Monteagudo

Los MASC como derecho humano para optar por otra forma de justicia y la IA como vía para facilitar su efectividad
Ana María Vall Rius

La inocencia de la responsabilidad social corporativa para proteger los derechos humanos ante la inteligencia artificial
Raúl López González

Impacto de la inteligencia artificial en los derechos de los interesados: una perspectiva práctica
María Luisa González Tapia

Desafíos ético-jurídicos en el uso de Inteligencia Artificial para el tratamiento masivo de datos biométricos
Nuria Cuadrado Gamarra

II. BOOK REVIEWS / CRÍTICAS BIBLIOGRÁFICAS

Instituto de Derechos Humanos Pedro Arrupe Human Rights Institute

Deusto Journal of Human Rights

Revista Deusto de Derechos Humanos

DOI: https://doi.org/10.18543/djhr

Deusto Journal of Human Rights is included in:
La *Revista Deusto de Derechos Humanos* está incluida en:

Deusto Journal of Human Rights
Revista Deusto de Derechos Humanos

No. 14

2024

DOI: https://doi.org/10.18543/djhr142024

Editorial Office / Oficina Editorial

Trinidad L. Vicente (Editor), Deusto Journal of Human Rights
University of Deusto
Pedro Arrupe Human Rights Institute
Apartado 1
48080 Bilbao, SPAIN
E-mail: revista.derechos.humanos@deusto.es
URL: http://djhr.revistas.deusto.es/

Apartado 1 - 48080 Bilbao, ESPAÑA
e-mail: publicaciones@deusto.es
Web: http://www.deusto-publicaciones.es/

ISSN: 2530-4275
ISSN-e: 2603-6002
Depósito legal: BI - 1.859-2016

Printed in Spain/Impreso en España

Deusto Journal of Human Rights

Revista Deusto de Derechos Humanos

No. 14/2024

DOI: https://doi.org/10.18543/djhr142024

Inteligencia Artificial y Derechos Humanos: desafíos y oportunidades en la era digital

Contents / Índice

Introducción al monográfico
José Miguel Iturmendi Rubia 11

I. ARTICLES / ARTÍCULOS

Human rights, vulnerability and artificial intelligence: an analysis in constitutional perspective
Jorge Castellanos Claramunt 33

AI in supply chains: freedom from slavery revisited
Migle Laukyte & Lorena María Arismendy Mengual 51

The systematics of the European Artificial Intelligence Act in the context of the fundamental rights of the Union: the myth of the digital constitutionalism
Ainhoa Lasa López 73

Facing fundamental rights in the age of preventive *ex ante* AI: a contemporary form of discrimination
Mª Teresa García-Berrio Hernández 101

Opportunities and challenges of AI chatbots for digital youth information, advice, and counselling services in Europe
Alonso Escamilla & Paula Gonzalo 127

The human right to participate and its connection to artificial intelligence
María Dolores Montero Caro 155

Towards a better protection of human rights through the use of AI and related technologies in budgeting and auditing of public expenditure
María Amparo Grau Ruiz 173

Ética, desafíos y riesgos del acceso a la justicia algorítmica
José Carlos Fernández Rozas 203

Ética en crisis: El impacto de la carrera armamentística de las armas autónomas en nuestros valores morales
Jorge Couceiro Monteagudo 237

Los MASC como derecho humano para optar por otra forma de justicia y la IA como vía para facilitar su efectividad
Ana María Vall Rius 259

La inocencia de la responsabilidad social corporativa para proteger los derechos humanos ante la inteligencia artificial
Raúl López González 287

Impacto de la inteligencia artificial en los derechos de los interesados: una perspectiva práctica
María Luisa González Tapia 313

Desafíos ético-jurídicos en el uso de Inteligencia Artificial para el tratamiento masivo de datos biométricos
Nuria Cuadrado Gamarra 341

II. BOOK REVIEWS / CRÍTICAS BIBLIOGRÁFICAS

Balcerzak, Michal and Julia Kapelańska-Pręgowska, eds. 2024. *Artificial Intelligence and International Human Rights Law. Developing Standards for a Changing World*. Cheltenham: Edward Elgar. 347 p. 377

Cotino, Lorenzo y Jorge Castellanos, eds. 2023. *Algoritmos abiertos y que no discriminen en el sector público.* Valencia: Tirant lo Blanch. 292 p. 385

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 9-10
http://djhr.revistas.deusto.es/

Inteligencia Artificial y Derechos Humanos: desafíos y oportunidades en la era digital. Introducción al monográfico

Artificial Intelligence and Human Rights: challenges and opportunities in the digital age. Introduction to the monograph

José Miguel Iturmendi Rubia
CUNEF Universidad. España
jmiturmendi@cunef.edu
ORCID: https://orcid.org/0009-0008-6402-9386

https://doi.org/10.18543/djhr.3202 Fecha de publicación en línea: diciembre de 2024

Cómo citar / Citation: Iturmendi, José Miguel. 2024. «Inteligencia Artificial y Derechos Humanos. Desafíos y oportunidades en la era digital». *Deusto Journal of Human Rights*, n. 14: 11-31. https://doi.org/10.18543/djhr.3202

Sumario: 1. Tecnología, sociedad y derecho. 2. El impacto de la IA en los derechos humanos. Recomendaciones. 3. En torno al monográfico. La incidencia de la inteligencia artificial en los derechos humanos. 4. Agradecimientos. Referencias bibliográficas.

1. **Tecnología, sociedad y derecho**[1]

Hace más de un siglo, el veinticuatro de enero de 1902, en el curso de una celebrada conferencia pronunciada en la *Royal Institution of London* que lleva por título "Descubrimiento del futuro", el novelista, reformador activo y lúcido polemista inglés Herbert George Wells pudo decir, desde su apasionada creencia en la perfectibilidad ilimitada del género humano por medio del progreso científico: "En el siglo pasado se

[1] La edición de esta obra, así como las páginas que siguen son parte de las actividades del Proyecto de I+D+i PID2022-136439OB-I00/ MCIN/AEI/10.13039/501100011033, Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas, financiado por el Ministerio de Ciencia e Innovación, cofinanciado por el Fondo Europeo de Desarrollo Regional "Una manera de hacer Europa", de cuyo Grupo de Trabajo formo parte. El investigador principal es Lorenzo Cotino Hueso.

produjeron más cambios que en los mil años precedentes […] pero los que habrán de verse en el nuevo siglo harán que aquellos nos parezcan pequeños" (Wells 1913, 62). ¿Qué no decir aquí y ahora, en el tercer milenio, con tantas y tan hondas transformaciones por obra de las tecnociencias? La irrupción de la Red contribuyó hace décadas a una expansión enorme de nuestra capacidad de comunicación inmediata a distancia a costes extremadamente bajos superando con éxito los estrictos condicionamientos de tiempo y espacio a los que se habían visto sometido los instrumentos de que entonces disponíamos. La inteligencia artificial (en adelante IA), instrumento dotado de una compleja serie de potencialidades, permite tareas de realización imposible en un pasado no muy lejano. Ingenios emblemáticos de nuestra cultura y de nuestra actual sociedad tecnológica que están llamados a constituirse en elementos determinantes de la futura memoria histórica y confirmarían la existencia presente de un Prometeo definitivamente desencadenado, "al que la ciencia le proporciona fuerzas nunca antes conocidas y la economía un infatigable impulso" (Jonas[2] 1995, 15).

Los hechos confirman con holgura la validez de la afirmación de Wells, aunque constituye una verdad indisputada que la experiencia del excepcional nivel de desarrollo tecnológico se caracteriza por una velocidad de crecimiento desconocida hasta hace bien poco. Resulta palmario que la tecnología "ha apretado a fondo el acelerador" (Cuadrado 2020, 1650), revolucionando las relaciones organizativas y renovando de manera radical los modos de producción. Asimismo, el grupo de usuarios de los avances tecnológicos, que ha pasado a configurar un conjunto harto heterogéneo, entre los que se cuentan los más variados organismos, instituciones, empresas, profesionales y consumidores, deben adaptarse a los cambios. La denominada "cuarta revolución industrial"[3] (Schwab 2016) ha generado un proceso transformador que quizás concluya por ser imparable en tantos y en

[2] La obra filosófica escrita por Han Jonas, *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*, tuvo un éxito de ventas insólito para las publicaciones de su naturaleza, y obtuvo una difusión tan amplia como rápida, con el reconocimiento académico al tiempo que mediático. En la misma se presenta una reformulación del imperativo categórico kantiano –"Obra de tal manera que las consecuencias de tu acción sean compatibles con la permanencia de una auténtica vida sobre la tierra". El filósofo alemán invoca una ética planetaria y expresa la preocupación por el futuro y la conveniencia de no bajar la guardia frente a los peligros y las amenazas de una técnica que, junto a sus aportaciones de progreso y bienestar, pretende someter y explotar la Naturaleza.

[3] Revolución o, como se ha señalado, "evolución, por cuanto perfeccionada de las herramientas desarrolladas por la tercera Revolución Industrial" (Lasa 2023).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 11-31
https://doi.org/10.18543/djhr.3202 • http://djhr.revistas.deusto.es/

tan variados ámbitos de la realidad y del conocimiento. Se trataría de una revolución continua (Fragonard 1995, 199-210), de ser posible la existencia de tal género de revoluciones, que ha convertido de este modo a las transformaciones de las tecnociencias en el paradigma conceptual dominante en nuestro tiempo.

Los avances tecnológicos, de distinta condición y continuamente susceptibles de renovación, y de hecho renovados sin solución de continuidad, han determinado, entre otras novedades, que sus posibles desarrollos, rumbo futuro y conclusión no parezcan de predicción fácil por los analistas; aunque no por ello estamos carentes de augures y escudriñadores del futuro posible. "Algún día hemos de llegar / luego sabremos donde", que se dice en el poema nacional argentino "El gaucho Martín Fierro" (Hernández 1872).

La IA ha terminado por convertirse en el sector estrella absoluto de las últimas semanas en los mercados de valores y conoce una verdadera explosión alcista. En la última década hemos sido testigos de la consolidación, el desarrollo y el éxito de un ingenio, la IA, cuya implementación en múltiples ámbitos de las actuales sociedades se produce en lapsus de tiempo que cada vez resultan ser mucho más cortos. Y todo ello se produce además con un acusado ritmo de mudanza, que nunca ha dejado de multiplicar su velocidad de emergencia y de desarrollo, lo que nos permite afirmar, sin que padezca la verdad, que se trata de uno de los indicadores de la definición del presente. De este modo, el resto de los restantes indicadores que caracterizan nuestro tiempo, en una visión probablemente no exenta de exageración, bien podrían ser reducidos a la condición de meros apéndices subalternos, las más de las veces mutuamente intercambiables o simples flecos, figurantes añadidos al rasgo tecnológico que ha terminado por convertirse en protagonista, acaparador indisputado de todos los más destacados títulos de crédito.

La revolución científica-técnica conforma en una medida tal nuestro ámbito cultural, que no resulta difícil advertir el desarrollo en este del fetichismo del "objeto tecnológico" que encierra su utilización en la fase actual de la evolución de los seres humanos, así como el efecto *totem* que de ordinario ejerce. Al margen del inmenso arsenal de las novedosas tecnologías, la era de la IA ha traído consigo su propia mitología, sus objetos y altares de culto, sus mitos y sus lenguajes[4], sus amenazas y esperanzas.

[4] El vocabulario no puede por menos que adaptarse al nuevo escenario, al punto que las nuevas palabras clave y emblemáticas de nuestra cultura y de nuestra actual

Estaríamos así ante una situación que se ha alcanzado, recurriendo para ello siempre, entre otros elementos, al apoyo y al refuerzo que nos proporcionan una serie de elementos retóricos que no son una mera envoltura en orden a la mejor presentación del contenido del discurso científico–técnico, sino, más bien por el contrario, una de sus partes constitutivas más esenciales. De esta manera, se afianzaron y desplegaron tres ideas–fuerza que contribuyeron a prestigiar al discurso científico–técnico, y a reforzar el ya de por sí elevado grado de consenso existente en torno al mismo: a) el progreso técnico debe ser considerado ética y moralmente neutro; b) la reglas tecnológicas pueden ser consideradas de una máxima racionalidad en la medida en que prescriben el curso de la acción humana óptima, y c) la tecnología está provista de una ilimitada capacidad en orden a la resolución de los problemas que pueda llegar a generar (Coolen 1987, 41-65). Al tiempo que se acrecienta la confiabilidad del conocimiento de los expertos y de la tecnología, de forma que la creciente conciencia del riesgo no perturba sin embargo en muchas ocasiones la conciencia cotidiana confiada.

Todo ello se compadece, a la vez que se encuentra alumbrado, por el principio de libertad de innovación y la libre extensión del conocimiento, que parece haber adquirido la condición de valor profundamente arraigado en nuestra sociedad, como un medio legítimo de cara a la realización de determinados fines de diversa índole -económicos, profesionales o individuales- y como una manera de aplicarnos a nuestros propios límites, actividad esta que nos define como especie, y que en el proyecto de los Modernos liberaría a los seres humanos a través del paulatino dominio de la Naturaleza. Tan es así que hace varias décadas se convirtió como políticamente incorrecta la tesis desplegada en el controvertido libro de Roger Shattuck (1996), en el que, en consideración a las nuevas realidades, sostuvo que tal vez haya llegado el momento de reanalizar críticamente el principio, tan básico desde la Ilustración, en cuya virtud el arte y la ciencia deberían gozar de libertad absoluta, no pudiéndose establecer trabas a su desarrollo. La libertad de innovación no es, empero, absoluta e irrestricta, y tiene sus límites en la seguridad y el respeto a los derechos esenciales.

Tras la implementación en nuestras sociedades de los sistemas de reconocimiento facial, asistentes virtuales, los algoritmos predictivos o

sociedad han pasado a ser: algoritmos, aprendizaje automático, trazabilidad, explicabilidad, sesgos, etc.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 11-31
https://doi.org/10.18543/djhr.3202 • http://djhr.revistas.deusto.es/

los automóviles autónomos, la IA está revolucionando cómo interactuamos con el mundo y entre nosotros. Sin embargo, este avance también plantea importantes preguntas sobre sus implicaciones éticas y su impacto en los derechos humanos. Así, los riesgos asociados al uso de la IA han provocado una transformación de algunas de las visiones del mundo, un cierto descrédito de la tecnología, y podemos encontrar críticas de muy distinto calado a los ingenios tecnológicos que habría conducido al tránsito "del ciberentusiasmo a la tecnopreocupación" (Innerarity 2020, 92) o a una "frankenfobia tecnológica" (Cortina 2024, 15). Se sostiene, asimismo, que nos hallaríamos inmersos en una "Infocracia" (Han 2022, 25) o en la tantas veces referenciada "era del capitalismo de la vigilancia" en la que se ha venido "a revitalizar el panoptismo digital" (García-Berrio 2023, 46) en el que lo sabuesos de la vigilancia husmean "la conducta en las profundidades de las personas […] porque [personas e investigadores] dejan a su paso un rastro de carne cruda y barata, clicable, que los capitalistas de la vigilancia gustosamente cazan y devoran […] las fieras no tardan en presentarse al festín" (Zuboff 2019, 373).

La potencia de los sistemas de IA a la hora de procesar datos ha terminado por generar una singular inquietud, recelo e incertidumbre y no son pocas las voces que alertan del peligro que implica la evolución de las técnicas hasta el punto de que las libertades más esenciales corren riesgos o pueden llegar a ser desterradas. Como hemos comentado en otros escritos, uno de los signos distintivos de la época presente es que en ella el progreso tecnológico se encuentra ineludiblemente asociado a elecciones de corte ético. Por ello, parece necesario establecer un marco ético adecuado que se encuentre respaldado por una normativa clara, taxativa, que potencie el respeto de los derechos humanos y contribuya a construir un sistema normativo digital que tenga por objeto la protección de la dignidad de la persona y los derechos humanos.

A este respecto es preciso señalar que nos encontramos en un renovado escenario en donde la IA ha pasado a ser un elemento esencial de una civilización. Un mundo que requiere la participación a la hora de establecer sus reglas tanto de instancias públicas de diferente ámbito territorial, transnacional, estatal y local, como de organizaciones privadas de muy diversa naturaleza. Los principales marcos regulatorios —la Unión Europea, China y Estados Unidos[5]—

[5] Véase Bradford (2024).

ofrecen diferencias que deben ser consignadas. Europa ha adoptado un enfoque distintivo en la regulación de la IA, marcando un contraste notable con otras regiones como Estados Unidos o China[6]. A través del Reglamento de Inteligencia Artificial (en adelante RIA)[7] la Unión Europea ha decidido no solo regular la tecnología, sino establecer un marco normativo proactivo que busca mitigar los riesgos antes de que ocurran, priorizando la protección de los derechos fundamentales y la seguridad de los ciudadanos. Por su parte, el Consejo de Europa aprobó recientemente el Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho[8]. Texto que, al reconocer jurídicamente una serie de principios éticos a los que se atribuye, por fin, carácter normativo, incorpora "la lírica" a la "prosa del RIA" (Cotino 2023, 177). Este modelo europeo, que considera la IA como un instrumento que debe operar bajo un conjunto claro de reglas que garantice su uso ético, se diferencia del enfoque de mercados como el estadounidense o el asiático, donde la innovación tecnológica se impulsa con mayor énfasis sobre la rapidez y la competitividad. En estos contextos, la regulación suele ser vista como un obstáculo que frena el progreso, y la actitud general es más reactiva, esperando a que los problemas surjan para luego corregirlos. De este modo, Europa no solo está diseñando políticas que regulen los desarrollos tecnológicos, sino que se propone anticiparse a las posibles vulneraciones de derechos, estableciendo medidas preventivas que aseguren una IA que sea segura, confiable y respetuosa con los principios democráticos, una postura que pone en relieve la prioridad de la protección frente a la simple aceleración de la innovación.

2. **El impacto de la IA en los derechos humanos. Recomendaciones**

La IA tiene el potencial de mejorar significativamente nuestras vidas en diferentes ámbitos y puede contribuir a generar las condiciones

[6] Véase Castellanos (2023).

[7] 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 y se modifican determinados actos legislativos de la Unión. Véase Simón y Cotino (2024)

[8] Decisión (EU) 2024/2218 del Consejo, de 28 de agosto de 2024, relativa a la firma, en nombre de la Unión Europea, del Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho. Publicado en: «DOUE» n. 2218, de 4 de septiembre de 2024: 1.3.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 11-31
https://doi.org/10.18543/djhr.3202 • http://djhr.revistas.deusto.es/

adecuadas que faciliten que los derechos y libertades de que gozan los ciudadanos puedan ser ejercidos de una forma más eficaz. Así, en el ámbito de la protección de la salud, la implementación de sistemas de IA permite diagnósticos más precisos, acelerar los mismo y tratamientos personalizados. En el entorno educativo ofrecen herramientas adaptativas que permiten atender las necesidades específicas de cada estudiante. En relación con el derecho a una tutela judicial efectiva, los algoritmos pueden optimizar el acceso y reducir las dilaciones indebidas. Sin embargo, también surgen preocupaciones sobre cómo estas tecnologías afectan derechos como la privacidad, la igualdad, la libertad de expresión y el debido proceso. Si bien estas tecnologías pueden mejorar la eficiencia y la imparcialidad en ciertos procedimientos, también existe el riesgo de que decisiones automatizadas comprometan el derecho a un juicio justo. La falta de transparencia en los algoritmos dificulta la posibilidad de apelación o revisión, erosionando principios básicos del Estado de Derecho (Cotino y Castellanos 2023).

Uno de los retos más significativos es el uso masivo de datos personales para entrenar sistemas de IA. Los datos biométricos, como huellas dactilares, reconocimiento facial o escaneo del iris son herramientas fundamentales en muchas aplicaciones tecnológicas. Aunque su uso puede incrementar la seguridad y eficiencia en diversos sectores, también plantea riesgos significativos para la privacidad y la protección de los derechos individuales. La recopilación y almacenamiento masivo de estos datos aumenta la vulnerabilidad frente a abusos, violaciones de la privacidad y usos discriminatorios.

Además, los sistemas de IA están íntimamente ligados a la problemática de los sesgos algorítmicos. Estos pueden surgir de los datos con los que se entrenan los sistemas o de decisiones tomadas durante su diseño. Los sesgos algorítmicos pueden perpetuar y amplificar desigualdades existentes, afectando desproporcionadamente a grupos vulnerables como mujeres, minorías étnicas, personas con discapacidad y comunidades marginadas. Esto pone en riesgo derechos fundamentales como la igualdad e interdicción de la discriminación.

La libertad de pensamiento, conciencia y religión también puede verse amenazada por el uso indebido de estas tecnologías. Los sistemas de IA pueden manipular preferencias y comportamientos a través de técnicas de personalización, lo que podría derivar en un control sutil pero poderoso sobre las creencias y elecciones de las personas. Por ello, garantizar la transparencia y la explicabilidad en los sistemas de IA es crucial para mantener la autonomía de los usuarios y proteger estos derechos.

Por otro lado, la relación entre la IA y los derechos de los menores es un tema especialmente delicado. Los niños y adolescentes son particularmente vulnerables al impacto de las tecnologías digitales, incluyendo la IA. Desde el uso de algoritmos en plataformas de redes sociales hasta la incorporación de IA en herramientas educativas, las interacciones de los menores con estas tecnologías pueden influir en su desarrollo cognitivo, emocional y social. Uno de los principales riesgos radica en la exposición a contenidos inapropiados o manipuladores y en la recopilación masiva de datos personales sin un consentimiento adecuado. Por ello, es fundamental que las aplicaciones de IA destinadas a menores cumplan con estrictos estándares éticos y legales, priorizando siempre el interés superior del niño y garantizando un entorno seguro y respetuoso de sus derechos fundamentales.

Asimismo, el derecho a la protección de la salud es uno de los ámbitos donde la IA puede tener un impacto transformador. Los sistemas de IA permiten avances significativos en la medicina personalizada, facilitando diagnósticos más rápidos y precisos, así como tratamientos adaptados a las necesidades específicas de cada paciente. Además, la IA ha demostrado ser crucial en la gestión de crisis sanitarias, como en la predicción y control de pandemias. Sin embargo, su implementación no está exenta de retos. La posible discriminación en el acceso a tecnologías médicas avanzadas, la privacidad de los datos de salud y la dependencia excesiva de sistemas automatizados son aspectos que requieren una regulación cuidadosa. Es esencial garantizar que la IA se utilice para fortalecer los sistemas de salud y que su desarrollo se alinee con los principios de equidad, accesibilidad y respeto por los derechos humanos.

Un aspecto crucial en el desarrollo y uso de la inteligencia artificial es la gobernanza, entendida como el conjunto de políticas, normativas y prácticas destinadas a guiar su implementación responsable. La gobernanza de la IA debe asegurar que estas tecnologías se utilicen en beneficio de la sociedad, estableciendo controles para mitigar riesgos y promover el cumplimiento de valores éticos y democráticos. Esto incluye la creación de marcos regulatorios que garanticen la transparencia, la rendición de cuentas y la participación ciudadana en la toma de decisiones relacionadas con la IA.

En este contexto, el fenómeno de las *fake news* adquiere una dimensión especial. Los algoritmos de IA tienen la capacidad de difundir información falsa a una escala y velocidad sin precedentes, afectando procesos democráticos y socavando la confianza en las instituciones. Los sistemas diseñados para identificar y combatir la desinformación se enfrentan al desafío de equilibrar la libertad de

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 11-31
https://doi.org/10.18543/djhr.3202 • http://djhr.revistas.deusto.es/

expresión con la necesidad de proteger la integridad del debate público. La gobernanza debe, por tanto, promover el desarrollo de tecnologías que refuercen la veracidad y la calidad de la información sin vulnerar derechos fundamentales.

Los chatbots, como herramientas basadas en IA, representan una innovación disruptiva en la comunicación y el acceso a servicios. Desde su uso en atención al cliente hasta su integración en servicios públicos, los chatbots pueden mejorar significativamente la eficiencia y accesibilidad. Sin embargo, su implementación plantea preguntas sobre privacidad, transparencia y calidad en la interacción. En el ámbito de los derechos humanos, los chatbots pueden ser utilizados para proporcionar asesoramiento jurídico, apoyo psicológico o información sobre derechos fundamentales, especialmente en contextos donde el acceso a estos servicios está limitado. No obstante, es esencial garantizar que estos sistemas sean inclusivos, libres de sesgos y diseñados para proteger los datos sensibles de los usuarios. Además, debe considerarse el impacto en el empleo y la relación humana, ya que la automatización excesiva puede deshumanizar ciertos procesos.

En este contexto, la regulación es fundamental. La Unión Europea ha liderado la discusión con propuestas como el Reglamento Europeo de IA, que busca establecer un marco para la gestión de riesgos y categoriza los sistemas de IA según su nivel de impacto. Este enfoque preventivo permite abordar problemas éticos y legales antes de que ocurran, protegiendo los derechos fundamentales sin sofocar la innovación.

No obstante, la regulación por sí sola no es suficiente. Es crucial fomentar una colaboración interdisciplinaria entre gobiernos, sector privado, academia y sociedad civil. La inclusión de voces diversas en el diseño y la implementación de sistemas de IA asegura que estas tecnologías sean inclusivas, éticas y responsables. También es esencial promover una alfabetización digital[9] que permita a los ciudadanos

[9] El Considerando 20 del RIA Considerando (20), establece que "la alfabetización en materia de inteligencia artificial debe proporcionar a todos los agentes pertinentes de la cadena de valor de la IA los conocimientos necesarios para garantizar el cumplimiento adecuado y la correcta ejecución. Además, la puesta en práctica general de medidas de alfabetización en materia de IA y la introducción de acciones de seguimiento adecuadas podrían contribuir a mejorar las condiciones de trabajo y, en última instancia, sostener la consolidación y la senda de innovación de una IA fiable en la Unión". Y, además, en el artículo 4 del RIA podemos leer que "los proveedores e implantadores de sistemas de IA adoptarán medidas para garantizar, en la medida de lo posible, un nivel suficiente de conocimientos de IA de su personal y de otras personas que se ocupen del funcionamiento y uso de los sistemas de IA en su nombre, teniendo

comprender los riesgos y beneficios de la IA, así como ejercer un control informado sobre su uso.

Las aplicaciones de la IA en el ámbito laboral también son motivo de preocupación y análisis. Los algoritmos que evalúan el desempeño en tareas laborales sirven para seleccionar personal o distribuyen diferentes tareas entre los trabajadores, y tienen un impacto directo en la vida de los mismos. Si no se gestionan adecuadamente, pueden generar discriminación o aumentar la precariedad laboral, o mermar los derechos laborales (Mercader 2022). Es esencial que se establezcan directrices claras para garantizar un uso justo y ético de estas herramientas, protegiendo el derecho al empleo digno.

Asimismo, el impacto de la IA en la esfera medioambiental no debe ser subestimado. Aunque esta tecnología puede contribuir a la sostenibilidad, optimizando procesos y reduciendo el desperdicio de recursos, también genera una huella de carbono significativa debido al alto consumo energético de los centros de datos y el entrenamiento de algoritmos complejos. Este aspecto plantea la necesidad de diseñar soluciones tecnológicas que sean no solo eficaces, sino también sostenibles y respetuosas con el medio ambiente.

Por otro lado, en el ámbito educativo, la IA ofrece oportunidades sin precedentes para personalizar el aprendizaje y ampliar el acceso a la educación. Sin embargo, también existe el riesgo de que estas herramientas refuercen desigualdades preexistentes si no se garantiza un acceso equitativo. Es crucial que las políticas públicas promuevan la inclusión digital y la formación en competencias tecnológicas, asegurando que nadie quede atrás en esta transformación.

Podemos afirmar sin riesgo a equivocarnos que la inteligencia artificial representa una oportunidad única para el avance humano, pero también una responsabilidad colectiva para garantizar que su desarrollo respete y promueva los derechos fundamentales. El equilibrio entre la innovación tecnológica y la protección de valores democráticos es el desafío central de nuestra era. Al abordar estos retos con rigor y compromiso, podemos construir un futuro donde la IA sea una herramienta para el bienestar y la justicia global.

A modo de conclusión, podemos afirmar que la IA representa una herramienta de enorme potencial para mejorar diversos aspectos de la vida humana, desde la salud y la educación hasta la justicia y la sostenibilidad medioambiental. Sin embargo, su implementación

---

en cuenta sus conocimientos técnicos, experiencia, educación y formación y el contexto en el que vayan a utilizarse los sistemas de IA, y considerando las personas o grupos de personas sobre los que vayan a utilizarse los sistemas de IA".

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 11-31
https://doi.org/10.18543/djhr.3202 • http://djhr.revistas.deusto.es/

masiva plantea riesgos significativos para los derechos fundamentales. Entre estos riesgos destacan la discriminación derivada de sesgos algorítmicos, la vulneración de la privacidad mediante la recopilación y tratamiento de datos personales, y el debilitamiento de valores democráticos por la propagación de desinformación y el uso indebido de sistemas automatizados en la toma de decisiones. La gobernanza de la IA se perfila como un desafío central para garantizar que esta tecnología contribuya al bienestar social sin comprometer los derechos humanos. La participación activa de gobiernos, empresas, academia y sociedad civil es esencial para establecer un marco normativo que promueva la transparencia, la explicabilidad y la inclusión. Además, la alfabetización digital y el fortalecimiento de capacidades ciudadanas son herramientas clave para empoderar a los individuos frente al impacto de la IA. En ámbitos específicos como la protección de menores y el derecho a la salud, la IA ofrece oportunidades únicas, pero también exige un enfoque ético y regulaciones que prioricen el bienestar de las personas. La IA debe ser diseñada y aplicada de manera inclusiva, asegurando que beneficie a todos los sectores de la sociedad sin perpetuar desigualdades o vulnerar derechos.

Como recomendaciones *ad futurum*, podemos señalar las siguientes:

Primera. Fortalecer los marcos regulatorios: Diseñar normativas claras que promuevan la transparencia, la rendición de cuentas y el respeto por los derechos fundamentales en el desarrollo y uso de sistemas de IA. El futuro desarrollo y progreso tecnológico debe resultar acorde con el sistema de libertades característico de las sociedades democráticas avanzadas y resulta imperativo evitar que el inevitable desarrollo de la IA termine por generar un detrimento de las garantías jurídicas de las personas. Los avances tecnológicos deben ser adecuadamente controlados y puestos al servicio de la mejora de las condiciones de vida, para lo cual necesitamos un orden jurídico adecuado. Las tecnologías emergentes no deben en ningún caso acentuar renovadas formas de desigualdad, pérdida de derechos fundamentales o menoscabo de la dignidad.

Segunda. Fomentar la alfabetización digital: Implementar programas educativos que capaciten a la ciudadanía en el uso responsable de tecnologías basadas en IA, con énfasis en poblaciones vulnerables. Bastantes de las afirmaciones y no pocas de las conclusiones que encontramos en las obras que componen este monográfico, sólo pueden ser adecuadamente comprendidas desde la asunción de una responsable consciencia tecnológica, tarea a la que debemos invitar a despertar y desarrollar. Actitud crítica, reflexiva y

responsable son cualidades necesarias ante los riesgos asociados a las tecnociencias y frente a los nuevos problemas.

Tercera. Garantizar la protección de datos: Reforzar las medidas para proteger los datos personales y limitar su uso a fines éticos y legales, evitando la explotación o manipulación de la información sensible.

Cuarta. Desarrollar tecnologías inclusivas: Promover la participación de comunidades diversas en el diseño de sistemas de IA para evitar sesgos y asegurar que estas herramientas reflejen las necesidades de todos los sectores de la sociedad.

Quinta. Priorizar la protección de menores: Establecer controles estrictos en plataformas digitales y sistemas de IA que interactúen con menores, garantizando su seguridad y bienestar.

Sexta. Optimizar el uso de la IA en salud: Asegurar que las aplicaciones de IA en el ámbito médico sean accesibles, equitativas y alineadas con los principios éticos, promoviendo el acceso universal a los avances tecnológicos.

Séptima. Combatir la desinformación: Desarrollar sistemas eficaces para identificar y mitigar la propagación de *fake news* mediante IA, preservando al mismo tiempo la libertad de expresión.

Octava. Incentivar la colaboración interdisciplinaria: Crear espacios de diálogo y cooperación entre expertos en tecnología, derecho, ética y política pública para abordar de manera integral los desafíos que plantea la IA.

Estas conclusiones y recomendaciones constituyen una hoja de ruta integral para maximizar los beneficios de la IA al tiempo que se mitigan sus riesgos, con el fin de garantizar un futuro en el que el progreso tecnológico sea plenamente compatible con la protección y promoción de los derechos humanos. En este contexto, la dignidad humana emerge como un principio rector fundamental, pues implica reconocer el valor intrínseco de cada persona y asegurar que el desarrollo tecnológico no comprometa, sino que refuerce, el respeto por este valor esencial. El RIA, en su considerando sexto, destaca esta perspectiva al afirmar: "Dadas las importantes repercusiones que la IA puede tener en la sociedad y la necesidad de generar confianza, es fundamental que la IA y su marco reglamentario se desarrollen de conformidad con los valores de la Unión consagrados en el artículo 2 del Tratado de la Unión Europea (TUE), los derechos y libertades fundamentales consagrados en los Tratados y, de conformidad con el artículo 6 del TUE, la Carta. Como requisito previo, la IA debe ser una tecnología centrada en el ser humano. Además, debe ser una herramienta para las personas y tener por objetivo último aumentar el bienestar humano".

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 11-31
https://doi.org/10.18543/djhr.3202 • http://djhr.revistas.deusto.es/

3. **En torno al monográfico. La incidencia de la IA en los derechos humanos**

El monográfico que presentamos en esta sede integra aportaciones que contemplan la intersección entre la IA y los derechos humanos desde una considerable variedad de perspectivas y con diversidad de enfoque con vistas a un diálogo jurídicamente interdisciplinario, de lo contrario no se asumirá la complejidad de la realidad de la que se ocupan los autores. Cierto es que, sin abordar toda su fértil problemática ni ofrecer un cuadro acabado que ambicione ser completo o exhaustivo de la temática, lo que hubiera excedido en mucho los propósitos con que se aborda la cuestión y, además, sería tarea imposible. Se trata, como es de sobra conocido, de uno de los sectores de mayor dinamismo de los ordenamientos jurídicos contemporáneos, y de un mundo, los retos que para los derechos humanos suponen los desarrollos ininterrumpidos de las tecnologías emergentes, en el que pocos datos pueden entenderse como definitivos.

A continuación, con este marco general establecido, realizaremos un somero repaso de las diferentes aportaciones individuales que los diversos autores han realizado sobre este tema.

El monográfico se abre con la aportación que lleva por título *Human rights, vulnerability and artificial intelligence: an analysis in constitutional perspective*, del profesor de Derecho Constitucional de la Universidad de Valencia, Jorge Castellanos, una de las voces más autorizadas en nuestra lengua a la hora de abordar la relación entre IA, democracia y derechos humanos. En su texto, Castellanos identifica, con oportunidad y acierto, el impacto de la IA sobre los derechos humanos desde una perspectiva constitucional, destacando cómo los avances tecnológicos pueden beneficiar, pero también amenazar, a los grupos más vulnerables. Castellanos subraya que la IA puede mejorar la eficiencia y resolución de problemas complejos, pero también perpetuar desigualdades y discriminar debido a sesgos algorítmicos. Asimismo, enfatiza la necesidad de garantizar que la tecnología no comprometa los derechos adquiridos, protegiendo especialmente a los menores, mujeres, migrantes y personas con discapacidad, quienes enfrentan riesgos desproporcionados. Propone, además, adaptar el progreso tecnológico a un marco ético que respete la dignidad humana y aborde los desafíos legales y sociales que plantea la IA. El autor concluye que solo una IA diseñada con enfoque en los derechos fundamentales puede contribuir a una sociedad más equitativa y democrática.

En el contexto de las cadenas de suministro globales, la IA emerge como una herramienta clave para abordar problemáticas profundas como la esclavitud moderna. El artículo *AI in supply chains: freedom from slavery revisited*, coescrito por Migle Laukyte, experta en Ética y Derecho digital y profesora en la Universidad Pompeu Fabra y Lorena María Arismendy, profesora de Derecho civil en CUNEF Universidad, plantea cómo la IA puede contribuir a identificar y combatir prácticas de trabajo forzoso y trata de personas en estas redes complejas. Las autoras destacan que, aunque la IA ha sido criticada por sus riesgos hacia los derechos humanos, también puede ser aprovechada para incrementar la transparencia en las cadenas de suministro mediante sistemas de vigilancia algorítmica y blockchain. Subrayan la necesidad de un desarrollo ético y regulado de estas tecnologías, para mitigar los sesgos y proteger los derechos fundamentales. Realizan un pertinente examen la idea de que la responsabilidad de combatir estas prácticas no recae solo en las empresas, sino en un marco colaborativo donde gobiernos, sociedad civil y tecnología trabajen conjuntamente.

La integración de la IA en la regulación europea abre nuevas fronteras en el debate sobre los derechos fundamentales y su interacción con el mercado digital. El capítulo *The systematics of the European Artificial Intelligence Act in the context of the fundamental rights of the Union: the myth of the digital constitutionalism* de la profesora de Derecho Constitucional en la Facultad de Derecho de la Universidad del País Vasco Ainhoa Lasa, examina con exhaustividad lo que sin duda constituye una de las cuestiones más representativas del nuevo escenario normativo, esto es, cómo el Reglamento Europeo de IA articula un marco jurídico diseñado para equilibrar los beneficios económicos de la tecnología con la protección de los derechos fundamentales. La profesora Lasa, en un texto complejo que acredita otra vez más su excelencia investigadora, destaca el riesgo de que la subjetividad digital derive en un modelo que priorice la lógica de mercado sobre valores como la igualdad y la dignidad humana. La autora cuestiona el denominado "constitucionalismo digital", planteando que la centralidad del mercado en el sistema europeo puede debilitar las garantías sociales y los derechos fundamentales. Concluye señalando que, para lograr que la IA contribuya verdaderamente al bienestar humano, debe existir un enfoque normativo que ponga en primer plano la protección de la persona frente a las dinámicas deshumanizadoras del capitalismo digital.

En un universo donde la IA redefine nuestras sociedades, el siguiente artículo: *Facing fundamental rights in the age of preventive ex ante AI: a contemporary form of discrimination,* firmado por la

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 11-31
https://doi.org/10.18543/djhr.3202 • http://djhr.revistas.deusto.es/

profesora Titular de Filosofía del Derecho en la Facultad de Derecho de la Universidad Complutense, Teresa García-Berrio Hernández, sitúa el foco en los desafíos éticos y legales que plantea su uso, especialmente en la protección de los derechos fundamentales, un asunto del que se ha ocupado en publicaciones anteriores. La autora encara, con el detalle y rigor que caracterizan la totalidad de su producción bibliográfica, el estudio de cómo el Reglamento Europeo de IA establece un marco normativo pionero para gestionar los riesgos asociados a esta tecnología, categorizando los sistemas de IA según su nivel de riesgo. García–Berrio destaca con acierto los peligros de los sesgos algorítmicos, que perpetúan la discriminación hacia grupos vulnerables, y propone principios éticos como la no maleficencia y la beneficencia para mitigar estas desigualdades. Además, subraya la necesidad de salvaguardar la dignidad humana frente a manipulaciones subliminales y vulnerabilidades explotadas por la IA. La autora pone de relieve que, en el contexto de la inteligencia artificial, resulta fundamental proteger la libertad de pensamiento, conciencia y religión frente a los riesgos de manipulación y sesgos algorítmicos, destacando que cualquier sistema que amenace estas libertades básicas debe ser rechazado para preservar la autonomía y dignidad de las personas. Sólo un marco ético compartido puede garantizar un uso justo y humano de esta tecnología, promoviendo valores como la empatía y la transparencia.

En un mundo cada vez más digitalizado, la IA redefine la forma en que los jóvenes acceden a servicios esenciales, y los chatbots se posicionan como herramientas clave en este cambio. El capítulo de Alonso Escamilla y Paula Gonzalo: *Opportunities and challenges of AI chatbots for digital youth information, advice, and counselling services in Europe*, explora el uso de chatbots en servicios digitales de información, asesoramiento y orientación juvenil en Europa, destacando su potencial para ampliar el alcance, mejorar la eficiencia y personalizar la atención. Sin embargo, también subraya los riesgos asociados, como la falta de transparencia, los sesgos algorítmicos y la posible exclusión de grupos vulnerables. Los debutantes enfatizan la importancia de co-diseñar estas herramientas junto con organizaciones juveniles para garantizar que se adapten a las necesidades reales de los jóvenes. Aseveran que la incorporación ética de chatbots no solo debe complementar, sino enriquecer, los servicios existentes para garantizar que estos avances tecnológicos respeten los derechos fundamentales de los usuarios más jóvenes.

La irrupción de la IA plantea nuevos desafíos y oportunidades en la construcción de democracias participativas, donde los derechos

fundamentales adquieren una dimensión clave. María Dolores Montero, profesora de Derecho Constitucional en la Facultad de Derecho de la Universidad de Córdoba, experta en el buen gobierno y en la relación entre los algoritmos, la democracia y los derechos fundamentales, analiza en el artículo que lleva por nombre *The human right to participate and its connection to Artificial Intelligence*, de una forma especialmente cuidadosa, el impacto que proyecta y extiende la expansión en curso de la IA en el derecho humano a participar en la esfera pública, subrayando tanto los riesgos como las posibilidades que esta tecnología ofrece para fortalecer o debilitar las instituciones democráticas. Por un lado, se ocupa de una cuestión fundamental al destacar que la IA puede facilitar el acceso a la información y mejorar la calidad del debate público mediante herramientas que procesen grandes cantidades de datos y permitan decisiones más informadas. Por otro, alerta sobre amenazas como la desinformación, la tribalización política y la creación de realidades paralelas que erosionan el tejido democrático. Montero insiste en que el desarrollo de la IA debe enmarcarse como un complemento de las capacidades humanas, no como un sustituto, asegurando que los valores democráticos y los derechos ciudadanos permanezcan en el centro del avance tecnológico.

Por otra parte, la catedrática de Derecho Financiero en la Facultad de Derecho de la Universidad Complutense de Madrid y *Visiting Professor of Law* en la *Northwestern University* , María Amparo Grau, aborda, en *Towards a better protection of human rights through the use of AI and related technologies in budgeting and auditing of public expenditure,* un asunto que ha constituido un punto central de referencia en algunas de sus publicaciones: el uso de la IA y tecnologías relacionadas en los procesos presupuestarios y de control del gasto público como una herramienta para optimizar la asignación de recursos y garantizar una mejor protección de los derechos humanos. Subraya, en una texto claro, conciso y completo, que la implementación de IA puede ayudar a prevenir irregularidades, reducir la corrupción y aumentar la eficiencia financiera, permitiendo un acceso más equitativo a servicios esenciales como salud y educación. Asimismo, destaca la necesidad de una supervisión eficaz para evitar que la tecnología comprometa los derechos fundamentales y asegura que la digitalización debe alinearse con objetivos de desarrollo sostenible, respetando principios éticos y legales. Grau enfatiza la importancia de establecer un equilibrio entre el uso de tecnologías avanzadas y la protección de los derechos humanos, proponiendo herramientas innovadoras como presupuestos basados en inteligencia artificial para

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 11-31
https://doi.org/10.18543/djhr.3202 • http://djhr.revistas.deusto.es/

mejorar la toma de decisiones y la transparencia en la administración pública.

La revolución tecnológica que supone la IA en el ámbito de la justicia abre un complejo debate sobre su impacto ético y legal y los desafíos para garantizar la igualdad, el acceso a la justicia y la transparencia. José Carlos Fernández Rozas, catedrático de Derecho Internacional Privado de la Facultad de Derecho de la Universidad Complutense de Madrid y miembro del *Institut de Droit International*, reflexiona en su valioso estudio *Ética, desafíos y riesgos del acceso a la justicia algorítmica* acerca del tema a que compromete su título, en el estilo claro y terso que es marca de la casa, sobre la justicia algorítmica, destacando cómo la IA puede transformar el acceso a la justicia, agilizando procesos y mejorando la eficiencia en la resolución de conflictos. En su aportación no dejan de destacarse importantes riesgos como la opacidad de los algoritmos, la perpetuación de sesgos y la erosión del derecho a un juicio justo. A través de casos emblemáticos como *State v. Loomis*, el autor analiza cómo la falta de transparencia y las desigualdades en el uso de estas herramientas tecnológicas pueden socavar los principios fundamentales del Estado de Derecho. En un sistema en constante digitalización, Fernández Rozas enfatiza la necesidad de marcos regulatorios sólidos, transparencia en el diseño de algoritmos y una colaboración ética entre tecnología y derecho para asegurar que la IA refuerce, en lugar de poner en peligro, los valores democráticos y los derechos humanos.

En el acelerado desarrollo de las armas autónomas, el uso de la IA redefine la moralidad y la ética de los conflictos armados, planteando serias cuestiones sobre el futuro de la violencia sistemática. Jorge Couceiro aborda cómo las Armas Autónomas Letales (LAWS) generan una tecnificación que diluye la responsabilidad moral y deshumaniza tanto a víctimas como a perpetradores. Al sustituir decisiones humanas por procesos algorítmicos se debilitan las restricciones éticas tradicionales en el uso de la fuerza militar. Aunque sus defensores destacan la eficiencia y precisión de estas tecnologías, Couceiro destaca que perpetúan patrones históricos de violencia fría y calculada, socavando principios esenciales del Derecho Internacional Humanitario. Frente a estos dilemas, el debutante subraya la necesidad de un enfoque ético y regulatorio que priorice la dignidad humana por encima de los beneficios estratégicos inmediatos.

La combinación de métodos de resolución de conflictos y el uso de IA redefine la justicia al enfatizar la participación activa de las personas implicadas, aunque no está exenta de desafíos éticos. Ana María Vall, experta en mediación y profesora de Historia del Derecho en CUNEF

Universidad, destaca que los Métodos Adecuados de Solución de Controversias (MASC) representan una evolución hacia una justicia más colaborativa y humanizada, con la IA como aliada clave para superar barreras logísticas y ampliar su alcance. Sin embargo, la autora también señala que confiar en exceso en sistemas tecnológicos puede deshumanizar procesos profundamente emocionales y complejos, planteando dudas sobre la capacidad de la IA para captar la riqueza de las interacciones humanas y las dimensiones éticas inherentes a los conflictos. Este enfoque subraya la necesidad de un desarrollo ético de estas tecnologías para asegurar que complementen, en lugar de sustituir, la experiencia humana en la búsqueda de soluciones justas y personalizadas.

El vínculo entre la ética y la responsabilidad social corporativa (RSC) es esencial para garantizar que la IA se utilice de manera compatible con los derechos humanos fundamentales. Raúl López González explora en su aportación cómo la RSC puede convertirse en un pilar para la protección de derechos como la privacidad y el acceso al trabajo en la era digital, destacando el papel de las empresas en la adopción de medidas éticas y responsables en el desarrollo y uso de tecnologías basadas en IA. Subraya la importancia de integrar principios éticos en la toma de decisiones corporativas, asegurando que los avances tecnológicos contribuyan al bienestar social y al respeto por los derechos fundamentales, convirtiendo a la IA en una herramienta para el progreso sostenible y equitativo.

La IA está transformando el tratamiento de datos personales, planteando retos y oportunidades para el ejercicio de los derechos de los interesados. María Luisa González Tapia, abogada experta en Derecho Digital y *Counsel* en Ramón y Cajal Abogados desde hace más de una década, analiza cómo la implementación de la IA impacta en los derechos contemplados en el Reglamento General de Protección de Datos (RGPD), con especial énfasis en el artículo veintidós sobre decisiones automatizadas. Destaca la importancia de rediseñar las políticas internas para garantizar el ejercicio de derechos como el acceso, la rectificación y la oposición, y resalta el papel clave de la transparencia y la supervisión humana en el uso de sistemas automatizados. Su trabajo ofrece una visión práctica sobre cómo adaptar las herramientas tecnológicas a los marcos legales, asegurando un equilibrio entre innovación y protección de los derechos fundamentales.

El avance de la IA en el tratamiento de datos biométricos redefine los retos éticos y jurídicos en la protección de los derechos fundamentales, evidenciando tanto sus beneficios como los riesgos

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 11-31
https://doi.org/10.18543/djhr.3202 • http://djhr.revistas.deusto.es/

asociados. Nuria Cuadrado, profesora de Filosofía del Derecho de la Universidad Complutense de Madrid, pionera en la materia en España, dado que se ha ocupado del Derecho Digital desde el último lustro del pasado siglo, tras su defensa y posterior publicación de su Tesis Doctoral sobre sistemas expertos jurídicos, examina con exhaustividad en su sólida aportación la capacidad transformadora de la IA en sistemas biométricos como el reconocimiento facial y el escaneo de iris, subrayando su impacto en áreas como la seguridad y la identificación personal. La profesora Cuadrado destaca, no puede ser de otra manera, preocupaciones críticas relacionadas con la privacidad, la transparencia y los sesgos algorítmicos que pueden perpetuar desigualdades y discriminar a ciertos colectivos. Mediante el análisis del caso de Worldcoin, el artículo ilustra cómo la recopilación masiva de datos biométricos plantea desafíos regulatorios y éticos que requieren soluciones urgentes. Este trabajo enfatiza la necesidad de marcos normativos sólidos y un compromiso ético claro para garantizar que los avances tecnológicos no comprometan los derechos fundamentales y promuevan un desarrollo inclusivo y respetuoso. En un mundo de condición multipolar, parece claro que se requiere la participación a la hora de establecer sus reglas tanto de instancias públicas de ámbito territorial, transnacional y estatal y local, como de organizaciones privadas de muy diversa naturaleza.

Finalmente, es importante señalar que buena parte de las aportaciones que configuran el monográfico se desarrollan en el marco del Proyecto de I+D+i PID2022-136439OB-I00/MCIN/AEI/10.13039/501100011033, titulado "Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas", financiado por el Ministerio de Ciencia e Innovación y cofinanciado por el Fondo Europeo de Desarrollo Regional bajo el lema "Una manera de hacer Europa" y cuyo investigador principal el Lorenzo Cotino. Asimismo, se enmarca en el proyecto internacional "Developing SustAI'nAbility" (FEI-EU-23-02) y en el proyecto "Biosurveillance through Artificial Intelligence (AI) in the post COVID era: Corporality, Identity and Fundamental Rights" (TED Code 2021-129975B-C21). Estos proyectos convergen en el propósito de explorar las implicaciones éticas, sociales y legales de la inteligencia artificial en contextos globales, proporcionando un marco interdisciplinario para el análisis de cómo la IA está transformando nuestra sociedad. A través de estas iniciativas, se abordan tanto las oportunidades inéditas que la IA ofrece como los retos significativos que plantea en la protección de los derechos fundamentales, subrayando la necesidad de un desarrollo tecnológico centrado en valores humanos y el bienestar colectivo.

4. **Agradecimientos**

Las últimas líneas servirán como vehículo para expresar una serie de agradecimientos. No me resisto a poner de manifiesto mi agradecimiento a todas las personas que han participado en este número por su rigurosa dedicación y por ocuparse de asuntos relacionados con un proceso transformador que se encuentra en una deriva cuya previsión resulta harto azarosa para los expertos que apenas pueden bucear más allá de la delgada capa de la superficie, en el profundo océano de los desarrollos aún no definidos. Se trata, como es de sobra conocido, de uno de los sectores más dinámicos de los ordenamientos jurídicos contemporáneos y de un mundo, los retos que para los derechos humanos suponen los desarrollos ininterrumpidos de las tecnologías emergentes, en el que pocos datos pueden entenderse como definitivos. Su esfuerzo por abordar los complejos desafíos éticos, jurídicos y sociales que plantea la IA ha enriquecido enormemente el debate y nos impulsa a seguir reflexionando sobre el futuro de nuestras sociedades en este nuevo paradigma tecnológico.

Asimismo, extiendo mi gratitud a la Revista Deusto de Derechos Humanos por crear un espacio académico de excelencia para la divulgación y discusión de temas tan relevantes y necesarios en el contexto actual. De manera especial, quiero reconocer la labor de su directora, Trinidad L. Vicente por su dedicación y compromiso en la promoción de los derechos humanos desde una perspectiva multidisciplinar e innovadora. Su trabajo incansable hace posible que estas iniciativas sigan contribuyendo al avance del conocimiento y a la construcción de una sociedad más justa y equitativa. Un importante estímulo a la hora de aceptar la colaboración en la edición de este monográfico me lo proporcionó, vaya por delante, su inconfundible actitud personal, sus positivas maneras y sus modos característicos de comportarse frente a las más variadas exigencias y retos.

Gracias a todos por su valiosa aportación.

Referencias bibliográficas

Bradford, Anu. 2024. *Imperios digitales. La batalla global por la tecnología que marcará la geopolítica del futuro*. Barcelona: Shackelton books.

Castellanos, Jorge. 2023. «Sobre los desafíos constitucionales ante el avance de la Inteligencia Artificial. Una perspectiva nacional y comparada». *Revista de Derecho Político* 118: 261-87. Doi.org/10.5944/rdp.118.2023.39105.

Cotino, Lorenzo y Jorge Castellanos. 2023. *Algoritmos abiertos y que no discriminen en el sector público.* Valencia: Tirant lo Blanch.

Cotino, Lorenzo. 2023. «El Convenio sobre inteligencia artificial, derechos humanos, democracia y Estado de Derecho del Consejo de Europa». *Revista Administración y Ciudadanía* 18: 161-182. Acceso el 2 de diciembre de 2024: https://egap.xunta.gal/revistas/AC/article/view/5173/9431.

Cortina, Adela. 2024. *¿Ética o ideología de la inteligencia artificial? El eclipse de la razón comunicativa en una sociedad tecnologizada.* Barcelona: Paidós.

Cuadrado, Nuria. 2020. «Implicaciones ético–jurídicas de los sistemas de reconocimiento facial». En *Estudios jurídicos multidisciplinares*, dirigido por Mª José Falcón y Tella y Juan Antonio Martínez Muñoz (1647-1660). Valencia: Tirant lo Blanch.

García–Berrio, Teresa. 2023. «La sociedad digital como cultura del riesgo. Desafíos éticos y legales del uso de sistemas de inteligencia artificial para la evaluación de riesgos y la vigilancia preventiva». En *Inteligencia artificial y democracia: garantías, límites constitucionales y perspectiva ética ante la transformación digital,* dirigido por Jorge Castellanos (39-65). Barcelona: Atelier.

Fragonard, Michel. 1995. *Le culture du 20ème siècel: dicctionnaire d'histoire culturelle*. Paris: Bordad.

Han, Byung–Chul. 2022. *Infocracia. La digitalización y la crisis de la democracia*. Barcelona: Taurus.

Hernández, José. 1872. *El gaucho Martín Fierro.* Acceso el 4 de junio de 2024. https://www.cervantesvirtual.com/obra-visor/el-gaucho-martin-fierro--0/html/ff29f9cc-82b1-11df-acc7-002185ce6064_11.html.

Jonas, Hans. 1995. *El principio de responsabilidad. Ensayo de una ética para la civilización tecnológica*. Barcelona: Herder.

Innerarity, Daniel. 2020. «El impacto de la inteligencia artificial en la democracia». *Revista de las Cortes Generales* 109, 87-103. Doi:10.33426/rcg/2020/109/1526.

Lasa, Ainhoa. 2023. «La morfología jurídica del poder global de mercado en su fase digital: una reflexión constitucional». En *Inteligencia artificial y democracia: garantías, límites constitucionales y perspectiva ética ante la transformación digital,* dirigido por Jorge Castellanos (115-149). Barcelona: Atelier.

Mercader, Jesús R. 2022. *Algoritmos e inteligencia artificial en el derecho digital del trabajo.* Tirant lo Blanch: Valencia.

Presno, Miguel Ángel. 2023. *Derechos fundamentales e inteligencia artificial.* Madrid: Marcial Pons. Ediciones Jurídicas y Sociales.

Shattuck, Roger. 1996. *Conocimiento Prohibido*. Madrid: Taurus.

Simón, Pere y Lorenzo Cotino (dirs.). 2024. *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*. Madrid: Aranzadi.

Wells, Herbert George. 1913. *The Discovery of the Future.* New York: Huebsch.

Zuboff, Shoshana. 2019. *La era del capitalismo de la vigilancia. La lucha por un futuro humano frente a las nuevas fronteras del poder*. Barcelona: Paidós.

Human rights, vulnerability and artificial intelligence: an analysis in constitutional perspective

Derechos humanos, vulnerabilidad e inteligencia artificial: un análisis en perspectiva constitucional

Jorge Castellanos Claramunt
Universitat de València. Spain
jorge.castellanos@uv.es
ORCiD: https://orcid.org/0000-0001-9621-6831

https://doi.org/10.18543/djhr.3187

Submission date: 08.04.2024
Approval date: 16.09.2024
E-published: December 2024

Citation / Cómo citar: Castellanos, Jorge. 2024. «Human rights, vulnerability and artificial intelligence: an analysis in constitutional perspective.» *Deusto Journal of Human Rights*, n. 14: 33-50. https://doi.org/10.18543/djhr.3187

Summary: Introduction. 1. Human rights in technological terms. Special attention to vulnerability. 2. An assessment of areas of special protection for vulnerable people in relation to the use of AI. 3. Discrimination derived from AI biases. Conclusions. References.

Abstract: This article addresses the impact of artificial intelligence (AI) on human rights from a constitutional perspective, focusing on the vulnerability of certain groups in the face of technological advances. After an introduction contextualising the relevance of the topic, human rights in the technological context are examined, with a particular focus on the vulnerability of certain groups. An assessment is made of the areas of special protection for these people in relation to the use of AI, and discrimination arising from algorithmic biases is discussed. The conclusions highlight the need for legal research in the field of AI to focus on ensuring that technological progress does not undermine human rights acquired over time. The importance of protecting vulnerable groups, whose vital development may be disproportionately affected by the impact of AI, is emphasised. It identifies areas where the advancement of AI may generate adverse effects on citizens' rights, underlining the importance of adapting this technological progress to the protection of human rights. It also highlights the risk of algorithmic biases in the processing of personal data, highlighting the need to protect the privacy and data of individuals as fundamental elements to ensure an AI that respects human rights. It concludes that only an AI that

respects these rights can contribute to a more advanced and just society, based on democratic principles.

Keywords: human rights, vulnerability, artificial intelligence, constitutional law, democracy.

Resumen: Este artículo aborda el impacto de la inteligencia artificial (IA) en los derechos humanos desde una perspectiva constitucional, centrándose en la vulnerabilidad de ciertos grupos frente a los avances tecnológicos. Tras una introducción que contextualiza la relevancia del tema, se examinan los derechos humanos en el contexto tecnológico, con especial énfasis en la vulnerabilidad de ciertos grupos. Se realiza una evaluación de las áreas de protección especial para estas personas en relación con el empleo de la IA, y se discute la discriminación derivada de los sesgos algorítmicos. Las conclusiones destacan la necesidad de que la investigación jurídica en el campo de la IA se centre en garantizar que el avance tecnológico no menoscabe los derechos humanos adquiridos a lo largo del tiempo. Se enfatiza la importancia de proteger a los grupos vulnerables, cuyo desarrollo vital puede ser afectado de manera desproporcionada por el impacto de la IA. Se identifican áreas donde el avance de la IA puede generar efectos adversos en los derechos de los ciudadanos, subrayando la importancia de adaptar este progreso tecnológico a la protección de los derechos humanos. Igualmente se destaca el riesgo de sesgos algorítmicos en la tramitación de datos personales, resaltando la necesidad de proteger la privacidad y los datos de los individuos como elementos fundamentales para garantizar una IA respetuosa con los derechos humanos. Se concluye que solo una IA que respete estos derechos podrá contribuir a una sociedad más avanzada y justa, fundamentada en principios democráticos.

Palabras clave: derechos humanos, vulnerabilidad, inteligencia artificial, derecho constitucional, democracia.

Introduction[1]

One of the most salient issues in the subject of the expansion of AI is that globalisation and technological advances have become so intrinsic to humanity itself that they both benefit and threaten it. So the use of AI in almost every aspect of human life has a direct bearing on the way we orient the way we organise ourselves as societies and, by extension, on the protection and guarantee of the rights of the inhabitants within it. For obvious reasons, this protection and guarantee must be given with greater attention to those people who are most vulnerable and also to those rights that are non-negotiable in any human organisation: human rights.

The universality of human rights must be specific, overcoming false reductionism, identifying a human characteristic as essential for the whole of humanity (Ballesteros 2003). And as Fernández Ruiz-Gálvez (1999) reminds us, the very notion of human rights as a cultural and historical concept, as a prepositive ethical and legal regulatory ideal, has from its origins carried with it an aim of universality, a vocation of being rights ascribed to all human beings, the ownership of which corresponds to all. All of this being aware of the successive generations of rights that have been implemented throughout history (Fernández Ruiz-Gálvez 1996). Thus, by applying or introducing a technology whose nature is as expansive as that which sustains artificial intelligence itself, the most expansive human rights themselves are clearly going to be challenged. The relationship between these concepts is direct.

Against this background, if one studies the issue at hand in this paper in depth, it is easy to see that AI can improve people's lives in a variety of ways, depending of course on how it is used. Thus, AI can help solve complex problems and perform tasks more efficiently and accurately, saving time and resources and improving people's quality of life. For example, AI can be used to diagnose diseases faster and more accurately, or to develop safer and more sustainable transport technologies. In addition, AI can also help address social and environmental challenges, such as climate change and poverty, by analysing data and developing innovative solutions. However, it is

[1] This work has been carried out within the framework of the R&D&I Project PID2022-136439OB-I00/ MCIN/AEI/10.13039/501100011033, *Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas*, funded by the Ministry of Science and Innovation, Co-funded by European Regional Development Fund "A way to make Europe".

important to bear in mind that the impact of AI on people's lives will depend on how it is used and how the challenges and risks associated with its use are managed.

Naturally, law is largely shaped by the different overlapping social scenarios and, consequently, AI has a lot to say in its development and its impact on society, especially on those who do not enjoy a privileged position. If we approach the introduction of technology with the aspiration of making people's lives easier, it is certainly always welcome. The only issue to be considered in that regard will be how to persuade humanity that the indiscriminate use of technology may result in a reduced ability to carry out everyday tasks. But that topic, without detracting from it, is for another field of research.

The question to be highlighted will be aimed at studying the impact on vulnerable people, especially on the human rights of all, of the massive use and incorporation of artificial intelligence in our lives.

Strictly speaking, the use of algorithms determines the possibility of inputting a quasi-unlimited amount of data from which a behaviour, an action or a consequence can be inferred. In the legal order, it is as much as extrapolating a result from a huge amount of data that makes a response possible from a legal point of view. However, the very incidence of the human aspect of law implies that we must be wary of an uncritical surrender to technology. The fact is that not every decision sponsored by artificial intelligence is necessarily the best. It will be objective, for obvious reasons, since it is not directed by an individual, by a person. But the elements that have been taken into account to make a certain decision will come either from the data provided by a person or from the artificial intelligence's own learning capacity. In any case, the note of objectivity can be preached, but not that of infallibility. The technological entity can be wrong. It cannot make a bad extrapolation of data, because it does not conceive of error in its conceptualisations, but it can make an error because of the data it uses to make the decision.

This opens an interesting range of possibilities to reflect on, since many parameters come into play in the legal scenario when making a decision. The legality of an action implies a certain interpretation of the human reality, and to end up with an element of which the notes of humanity are not predicated can lead to minor, major or disastrous mistakes, depending on the intensity of the decision and the rights at stake. Thus, when decisions based on artificial intelligence interact with human rights, those whose greater protection must be established by the legal system by definition, the risk of affecting people's lives in their most sensitive issues is extremely high.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 33-50
https://doi.org/10.18543/djhr.3187 • http://djhr.revistas.deusto.es/

Nevertheless, we can indicate that the use of AI in democracy can have both positive and negative effects. On the one hand, AI can help improve the efficiency and transparency of democratic processes[2], for example through the use of big data analytics technologies to identify patterns and trends in electoral preferences, or through the use of *chatbots* and voice systems to facilitate access to information and public services. However, the use of AI in democracy is also likely to raise concerns related to privacy, security, and information manipulation. For example, the use of artificial intelligence technologies to collect and analyse personal data may put citizens' privacy at risk, and the use of AI to influence political decisions may affect the integrity and fairness of democratic processes. Overall, the impact of AI on democracy will depend on how it is used and how the challenges and risks associated with its use are managed.

As can easily be seen, this is not a dilemma of mere technological progress or not, but rather what is at stake is the fact that the impact of artificial intelligence on human rights in a concrete way, and of technological progress in general, can disrupt the basic elements of societies and, in so doing, dynamite them.

Hence, the main reflection should be on how to adapt artificial intelligence in harmony with human reality. Because technological progress is as uncontroversial as it is debatable whether the incorporation of artificial intelligence into all aspects of life does not entail an associated risk.

1. Human rights in technological terms. Special attention to vulnerability

There are two main rationales supporting the adoption of a human rights perspective in the context of technological progress: first, there is an intrinsic argument, which recognises that a human rights-based approach is the most appropriate from an ethical or legal perspective. Second, there is the instrumental argument, which recognises that a human rights-based approach leads to better and more and sustainable outcomes in terms of human progress.

In the same vein, it should be noted that the Vienna Declaration emphasises the responsibility of states and international organisations

[2] And in this regard, it should not be forgotten that the right to political participation is a human right recognised in Article 21 of the Universal Declaration of Human Rights (Castellanos 2020).

in creating an enabling environment to ensure the full enjoyment of human rights. This implies the elimination of all forms of human rights violations and their underlying causes, as well as overcoming obstacles to the realisation of these rights. Furthermore, it highlights that the proliferation of extreme poverty is a factor limiting the full and effective enjoyment of human rights. This correlation is relevant in the context of the expansion of artificial intelligence, as the most vulnerable people may face greater challenges in accessing and benefiting from emerging technologies due to socio-economic disparities. Protecting human rights therefore implies addressing inequities in the access and use of artificial intelligence to ensure that its development and deployment does not perpetuate the exclusion and marginalisation of the most disadvantaged groups.

Regardless of any technological evolution developed, it will be necessary to focus on certain groups that, due to social, historical or any other type of circumstances, do not have the capacity to raise their voices to correct or denounce certain effects of progress in the course of their lives. We are referring, of course, to those vulnerable groups that may suffer the negative consequences of certain technological advances (Castellanos 2024). In this respect, Ballesteros (1999) affirms that human beings are more radically vulnerable than self-sufficient, which is why the review of these issues responds to a collective need.

Thus, since the beginning of the so-called specification process of human rights, consisting of a greater determination and specification of the subjects of rights, there is no doubt that concern for the most vulnerable groups in society has fortunately been growing. Thus, on the one hand, in the international sphere, there has been a proliferation of Universal Declarations as well as Treaties and Conventions relating to those groups of people whose common characteristic is that, for various reasons, they find themselves in a disadvantaged social position, particularly unprotected and defenceless, making them especially vulnerable. These are groups of people in need of special protection or guardianship, even if the threat that makes this protection necessary varies. In this regard, mention should be made, among others, of the Declaration on the Rights of the Child of 20 November 1959, the International Convention on the Elimination of All Forms of Racial Discrimination of 21 December 1965, the Convention on the Elimination of All Forms of Discrimination against Women of 18 December 1979, the Convention on the Rights of the Child of 20 November 1989, the International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families of 18 December 1990, the Convention on the Rights of

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 33-50
https://doi.org/10.18543/djhr.3187 • http://djhr.revistas.deusto.es/

Persons with Disabilities of 13 December 2006. All these groups, which for different reasons, historical, cultural, economic, etc., are particularly vulnerable, are more likely to suffer discrimination, marginalisation or any other type of violation of their rights in the face of the emergence of new human needs or new forms of threats and attacks on equality, freedom, solidarity and security of people. And we cannot forget that the United Nations 2030 Agenda has as its motto "leave no one behind".

On the other hand, vulnerability has been the subject of profound reflection in recent decades in the fields of philosophy and law. Classical authors have devoted their efforts to delving into the inherent fragility of human beings and their natural interdependence. We should mention here Alasdair MacIntyre (2001), Amartya Sen (2000), Martha Nussbaum (2006) and Zygmunt Bauman (2017) or Bauman (2013), among others.

A vulnerability that has undoubtedly been accentuated in the first two decades of this new century due to the occurrence of a series of circumstances such as the succession of a series of economic, health, climatic and also humanitarian crises caused by the succession of armed conflicts such as the one in Syria or now in Ukraine or the Gaza Strip, which are causing an unprecedented exodus of refugees. To this situation must undoubtedly be added the economic and social transformations that are taking place due to technological changes and the spread of new technologies and artificial intelligence which, in addition to giving rise to the so-called revolution 4.0 revolution, are calling into question our form of social and legal organisation based on work as a central pillar and are having a dramatic impact on the most vulnerable groups. Authors such as Cassilli (2021), who wonders whether mankind will finally be able to get rid of work thanks to robots, have warned about this circumstance, as well as Zuboff (2020) analysing from Orwell's *1984* dystopia (2013) to *The Age of Surveillance Capitalism* or Yuval Noah Harari (2016) who considers that "AI will be able to collaborate in the creation of human jobs in other ways (new jobs) in parallel".

It can be concluded that the incorporation of AI has many benefits, but it is also necessary to consider that there are many challenges and risks associated with this technology. The main one is that of generating some kind of discrimination. Disrupting human rights by generating unequal treatment is one of the main elements to work on from a legal perspective. And vulnerable groups are those most likely to suffer some kind of harm caused by the development of AI (Saiz Garitaonandia 2023, Terrones 2023), so it is necessary to

establish the specific areas in which the citizen in question may be harmed (Ammerman 2022) and to generate the necessary means to mitigate and eliminate this potential discrimination or unfavourable treatment.

For these reasons, there are ethical concerns regarding the use of artificial intelligence, such as the perpetuation of discriminatory biases, inequality, the widening of the digital divide, social divisions, lack of transparency in algorithms, among others. This may even lead to imbalances between or within countries. However, steps are being taken to control these problems through ethical recommendations by UNESCO and recent EU regulations. It is important to remember that AI is a great opportunity for humanity, but its limits need to be regulated to ensure that it is used in a way that respects human rights. Rather than holding back progress and innovation in AI, these ethical concerns should be an impetus to research and develop technologies in an ethical and sustainable manner, in line with the sustainable development goals of the UN's 2030 AGENDA.

In this way, we find areas susceptible to undergo a profound transformation due to the implementation of Artificial Intelligence and this evolution may affect certain vulnerable groups with greater impact, generating some discriminatory situation. Generally speaking, we address the issue with regard to vulnerable groups such as, firstly, minors. The impact of new technologies on the development of the rights of minors must necessarily be addressed. Electronic devices are omnipresent in the lives of our minors, an omnipresence that has innumerable effects on their development, education, cognitive capacities, relational capacities, affective-sexual life, social life and on their self-awareness and self-perception. There is an urgent need to analyse both the positive and negative consequences of artificial intelligence in their lives.

We must also look at the issue of gender. Discrimination against women in the use of artificial intelligence is due to a number of reasons, one of the main ones being the lack of diversity in the AI development team, which can lead to unconscious biases in model design and training. In addition, datasets are often used that reflect existing inequalities in society, which can perpetuate these biases in model performance. This leads to addressing these problems through diversity in AI development teams and conscious curation of datasets. These problems are also extrapolable to the migrant population and to certain discriminatory effects based on race. The same is true for older people with the issue of new technologies and the generation gap. And, of course, the impact on disability is also significant (Valle 2023).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 33-50
https://doi.org/10.18543/djhr.3187 • http://djhr.revistas.deusto.es/

Artificial intelligence can be a great tool for improving the integration and inclusion of people with disabilities, but it could also become an insurmountable barrier for many of them. Unsuspected prospects of a full and independent life would open up for them if the development of these technologies is done with the participation and for the benefit of all citizens, which is why the sum of all potentialities and capacities is required.

2. An assessment of areas of special protection for vulnerable people in relation to the use of AI

Within this framework of thinking about the impact of AI on human rights, it is essential to propose comprehensive research on how artificial intelligence may affect society and individuals. This research must include a thorough analysis of the associated benefits and risks, as well as a realistic understanding of the technical and economic aspects. It is also crucial to take into account national and international regulations that may have an impact on these issues, as well as to study relevant public and private initiatives, and the current state of academic and industrial research in these areas and related standards. Hence, in the following, we will briefly outline a number of areas that would benefit from further study.

First, the impact of AI on jobs and employment should be studied. Artificial intelligence has the potential to have a major impact on employment as, on the one hand, it can help improve productivity and efficiency in a variety of industries, which can create new jobs and increase the demand for employees with specific skills in AI-related fields. It can also automate repetitive and dangerous tasks, improving safety on the job. But on the other hand, AI may also have a negative impact on employment due to the automation of certain jobs, which could reduce the demand for workers in certain areas (Acemoglu and Johnson 2023). This could lead to an increase in unemployment and increased competition for remaining jobs, affecting low-skilled occupations to a greater extent and, consequently, directly affecting the most vulnerable groups. There could also be inequality in access to AI-related jobs, which could widen economic and social gaps. Indeed, there is no shortage of threats to the most vulnerable populations in the employment sector from the use of AI.

The main thing to assess in this area, therefore, will be its possible effects on employment and to take measures to mitigate its negative impacts, especially on the most vulnerable population. This includes

investment in training and education for AI-related skills, as well as the creation of occupational safety programmes for workers affected by automation.

Another impact to highlight is on means of transport such as the autonomous car. AI in relation to transportation can progressively impact beneficial issues such as increased road safety as autonomous cars use advanced sensors and machine learning technologies to detect and avoid accidents, which could significantly reduce the number of traffic accidents and related fatalities. It also leads to greater traffic efficiency as autonomous cars can communicate with each other and with traffic systems to avoid congestion and optimise traffic flow.

We can even see greater accessibility for people with disabilities or the elderly as autonomous cars could allow these people to move around more independently and safely. To which we can add the issue of parking as autonomous cars can be programmed to find a parking spot and park autonomously, which could reduce the need to build expensive car parks. However, there are also some challenges and concerns related to the use of autonomous cars, such as the cost of developing and producing these vehicles, the security of data collected by sensors, and the potential for autonomous cars to cause job losses in the professional driver sector. These difficulties directly affect the most vulnerable populations in that they pose a barrier to entry to a service that clearly discriminates on the basis of economic capacity. It also generates a dependence on the control of the data provided, requiring a greater willingness to provide all types of information, also affecting the possibility of certain disadvantaged groups (migrant population) in terms of stabilising and processing the information. And, ultimately, it has an impact on the loss of employment, which results in greater difficulties for the most vulnerable population in terms of retraining. The study should focus on analysing the extent to which the most vulnerable people are affected by the introduction of the autonomous car.

Moreover, in the interests of improved efficiency, remote healthcare will also advance as AI enables remote patient monitoring and diagnosis, allowing patients to receive medical care regardless of their geographic location. However, there are also concerns about the impact of AI on healthcare, such as the potential for AI algorithms to perpetuate discrimination and inequality in access to healthcare, the privacy and security of medical data, and the potential for medical professionals to lose skills and knowledge due to automation.

The risks involved focus on affecting the most vulnerable populations as they do not improve current problems in terms of access to the healthcare system, so perpetuating health-related

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 33-50
https://doi.org/10.18543/djhr.3187 • http://djhr.revistas.deusto.es/

problems through the application of AI would increase inequalities and difficulties in healthcare for more vulnerable people. In addition, the tendency for AI to solve problems in practice in terms of efficiency would lead to treatment impact studies evaluating criteria that do not respond to a human rights-based treatment characteristic.

If the possibilities of cure and treatment are reduced to the percentage of efficiency of the process and the effects derived from it, the most vulnerable population will necessarily receive discriminatory treatment for the sake of a series of parameters to be managed by AI that would not respond to a human medical criterion. In this respect, this point is particularly significant in the development of AI in health in order to implement AI systems based scrupulously on respect for human dignity and rights.

Another area to highlight must necessarily be the impact of AI on education. AI can enable personalisation of learning by analysing data on student performance and adapting teaching content accordingly, providing a more personalised learning experience.

It also generates the possibility of developing virtual tutors as AI can be used to develop virtual tutors that can provide personalised feedback and guidance to students. AI can even be used to automatically assess student work, which can save time and resources for teachers. However, the digital divide, data privacy, and the automation of tasks that could replace teachers pose a major danger to the impact of AI in education.

It is no coincidence that one of the main SDGs addressed by UNESCO's own Recommendation on the Ethics of Artificial Intelligence (2021) is No. 4 "Quality education" together with No. 10, which refers to the "Reduction of inequalities". With regard to children, this is particularly relevant as applications of artificial intelligence in people's lives, in general, can have a negative impact on the cognitive development of young people and on the creation of inequality gaps between different groups. For example, studies suggest that excessive use of electronic devices can contribute to problems with attention, memory and academic performance in children and adolescents. In addition, limited access to technology and lack of digital skills can increase inequalities among migrants, older people and people with disabilities. It is important to take these factors into account when developing technology and education policies and programmes to ensure that all groups have the necessary skills and access to benefit from technologies and are not marginalised.

Another area to consider that affects people is the relationship of AI's impact on energy. In this area, AI can predict consumption and

therefore optimise its generation, reducing costs and improving efficiency. It can predict when maintenance is likely to be needed, which can help reduce downtime and improve availability, it is also possible to generate renewable energy, with solar and wind applications, by monitoring and controlling weather factors, as well as developing smart grids, which can improve the efficiency and security of energy supply by monitoring and automated control of demand. All of this is based on the analysis of data collected from sensors and devices connected to the energy grid, allowing for more informed decision-making on energy generation, distribution and consumption. However, all these implications come against the backdrop of the complexity and security of smart grids, and the privacy of customer data. In addition, the most vulnerable groups will be subject to decisions based on efficient energy production, disregarding the personal costs they incur and hindering access to energy. So not only will decisions have to be objectified by efficiency criteria, but dealing with the issue from the perspective of vulnerable groups will imply the mitigation and progressive elimination of those barriers that prevent equal access.

Another notable area is the impact of AI on logistics. By using AI on the work of trucks and drivers, costs can be reduced and efficiency improved due to the ability to plan delivery routes and optimise the use of resources. Demand for products and services can also be predicted, which can help companies to better plan their inventory and logistics. There is even the possibility to monitor the progress of deliveries and the position of vehicles in real time, allowing for greater transparency and better coordination. This goes hand in hand with progressive automation in order processing, which can help reduce errors and improve processing speed.

Vulnerable groups may be disadvantaged by the elimination of jobs in which they can develop due to the progressive automation of these jobs and also by not including the demand requirements of their products in the efficiency parameters. The economic difficulties that normally accompany groups that suffer discriminatory treatment may make it unattractive for artificial intelligence to collect their data to forecast product demand, so that the lower economic capacity will be aggravated by greater difficulties in the supply of products that they demand and, consequently, by higher prices, which will deepen the discriminatory and unfavourable treatment of vulnerable groups.

Given its strategic nature in countries such as ours, it is important to look at the impact of AI on tourism. AI-based virtual assistants can help travellers plan their trips, make reservations and answer questions,

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 33-50
https://doi.org/10.18543/djhr.3187 • http://djhr.revistas.deusto.es/

providing personalised recommendations: based on the analysis of travellers' data and providing personalised recommendations on destinations, accommodation and activities. Also significant is the impact in terms of machine translations and language recognition, which can facilitate communication for travellers. All of this is focused on the vast amounts of data collected on travellers and tourism trends, allowing for more informed decision-making on tourism planning and marketing strategy. Difficulties will come from data privacy and ethics in the use of technology, as well as the destruction of tourism-related jobs.

As the service sector is a favourable scenario for the integration of vulnerable groups into the labour market, the increase in technological possibilities related to tourism will put upward pressure on prices, undermining the ability of vulnerable groups to access it.

It is obviously fundamental for the protection of human rights to analyse the impact of AI on the environment, due to the direct implications that its protection has on the defence of people's rights and quality of life. We assume that society as a whole, including marginalised groups, and the environment have an active role to play in the whole process of developing and using artificial intelligence. It is necessary to promote sustainability and environmental responsibility, and to encourage research in this direction to ensure that the development of artificial intelligence is carried out in a sustainable way for future generations. Thus AI can be used to analyse large amounts of data collected on biodiversity and help scientists detect patterns and trends that can be used to improve wildlife and ecosystem conservation, as well as monitor the reduction of greenhouse gas emissions and improve energy efficiency.

Accurate climate predictions will help farmers to plan their crops and communities to prepare for extreme weather events, and AI is also relevant to optimise recycling and waste management processes, helping to reduce the amount of waste sent to landfills and increase the efficiency of recycling processes. In terms of negative impact, the massive use of AI may require a large number of computers and electronic devices, which will increase the greenhouse gas emissions associated with the production of these devices and their disposal. In addition, the use of AI in agriculture and fisheries may increase overfishing and deforestation, all of which have a greater impact on vulnerable groups in terms of employment, access to certain more environmentally friendly goods and services, and access to pollution-free areas.

Another scenario to reflect on is the impact of AI on the financial sector. AI-enabled analysis of large amounts of financial data, such as

bank transactions, stock prices and economic data, will help investors and banks make informed decisions. It will also help automate financial processes, such as loan approval, fraud detection and investment management, reducing costs and increasing efficiency. AI will also be used for personalised financial advice to customers, such as investment recommendations and savings plans, helping customers to make informed financial decisions. Such developments may significantly harm vulnerable groups insofar as the use of AI in the financial sector may amplify economic inequalities by giving investors with access to AI an advantage over investors without, and may also increase the risk of fraud and vulnerability to cyberattacks. In addition, the possibility of denial of loans and access to financial services because of the handling of people's data, or precisely because they belong to vulnerable groups, can be multiplied exponentially. The product of all this may be a clear advance in discrimination due to the use of AI.

And finally, in this non-exhaustive selection of AI impacts in the sphere of impact on people's rights, the impact of AI on access to housing must also be considered, because AI can contribute to exclusion and inequality in access to housing if it is used to automate decision-making in housing allocation, which could lead to increased discrimination and exclusion of vulnerable groups. In addition, AI can contribute to higher housing prices through the automation of property valuation, which could make it more difficult for people on low incomes to access housing. The conclusion of all this is the necessary conviction from the public sector to develop policies to favour access to housing for all people, regardless of their economic or social situation.

3. **Discrimination derived from AI biases**

Artificial intelligence can be biased and discriminate against people if the data and machine learning models used to train it contain biases or misinformation. As is well known, AI systems rely on data that they collect and analyse to "learn" and perform tasks autonomously, so if the data used to train an AI contains biased information or prejudices, it is possible that the system will also reproduce those biases and discriminate against certain people or groups. For example, if an AI system is trained with data that shows a higher number of crimes committed by people of a certain ethnic group, it is possible that the system will "learn" to discriminate against that particular group. It is important to bear in mind that biases in AI can have serious consequences in areas such as security, health and justice, so it is

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 33-50
https://doi.org/10.18543/djhr.3187 • http://djhr.revistas.deusto.es/

essential to ensure that the data and models used to train AI are objective and free of bias.

There are several methods for detecting biases in artificial intelligence. One of the most common is statistical analysis of the data used to train the AI, to identify possible biases in the data patterns and in the way the system processes and uses that information. Quality assessment techniques for machine learning models can also be used to measure the performance of the system on specific tasks and compare it with that of other systems or humans. Another way of detecting biases in AI is by conducting tests or experiments to evaluate the behaviour of the system in different situations and contexts, and to verify if there are differences in its performance depending on variables such as gender, race, age or sexual orientation of the people interacting with it. In general, it is important to bear in mind that the detection of bias in AI requires a systematic and rigorous approach, including the participation of experts in the field and the use of appropriate tools and techniques.

Algorithmic discrimination is a phenomenon that occurs when AI systems or algorithms used to make automated decisions indirectly or unconsciously discriminate against certain individuals or groups. This can occur when the data used to train the AI or to develop the algorithms contains biased information or prejudices, or when the machine learning models used reproduce or reinforce those biases. Algorithmic discrimination can have serious consequences in areas such as health, education, justice and employment, and can affect equity and social justice. It is therefore important to address algorithmic discrimination by identifying and correcting biases in the data and models used to train AI, as well as by implementing measures and regulations to ensure the responsible and ethical use of these technologies.

Moreover, the introduction of biases also generates discriminatory situations, so we not only address the issue from the results but also from the generation of the data to be studied in order to reach these inferences. This leads us to study how to obtain this data on personal actions in order to be able to analyse the information and provide a response from a technological point of view. Privacy and data protection will therefore be directly related to the impact of human rights on the results obtained.

Conclusions

Artificial intelligence has dominated most research in recent years. Its exponential development and the novelty of its appearance in practically

any field has aroused the research curiosity of a huge number of people. Regardless of this, the importance of legal research in this field is determined by the impact that unbridled technological progress may have on people's human rights. Thus, successive research should focus on ensuring that technological progress does not come at the cost of overturning rights that have been achieved by people for generations.

Most legal studies, such as this one, are set in this context of opposing the rapid advancement of technological possibilities to the guarantees of human rights. Thus, in the development of the paper we have emphasised the question of the relevance of AI in the development of the lives of particularly vulnerable groups, as they will be subject to impacts with less protection. It is the task of law to determine which groups are likely to be negatively impacted in their life development and to establish the appropriate guidelines to curb this incidence. Likewise, in order to carry out this task, we have established a series of areas, not exhaustively, but at least with a certain projection, in which the overwhelming advance of AI can generate adverse effects on citizens' rights. Undoubtedly, the positive advances are also significant, and this has been explained in each of the areas, so the solution is not to reduce technological progress, in any case, but to adapt this necessary progress to the care and protection of citizens' rights, particularly their human rights, and especially those who, due to circumstances of all kinds, are among vulnerable groups.

And the last scenario we have described is that of danger related not to the advance in itself, but to some specific characteristic of the advance of AI in the impact of human rights. In this respect, the biases that may occur in the processing of people's data take on particular prominence. Bearing in mind that all this technology is nourished by the data generated by citizens through their activity, privacy and data protection must play a leading role in this matter, precisely because of the widely held idea in this work that dignity and human rights must be the cornerstone of all AI development. Only an AI that respects human rights will be able to establish a more advanced and just society, protected by the democratic scenarios in which it develops. This is the pillar on which to build the whole unstoppable future society, in which AI will play an increasingly important role.

References

Acemoglu, Daron, and Simon Johnson. 2023. *Poder y progreso. Nuestra lucha milenaria por la tecnología y la prosperidad*. Bilbao: Deusto.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 33-50
https://doi.org/10.18543/djhr.3187 • http://djhr.revistas.deusto.es/

Ammerman, Julia. 2022. «Las personas vulnerables ante el derecho a la protección de datos personales». In *La privacidad en el metaverso, la inteligencia artificial y el big data: Protección de datos y derecho al honor*, coordinated by Ángel Acedo Penco, 49-63. Madrid: Dykinson.

Bauman, Zygmunt. 2013. *Vidas desperdiciadas: la modernidad y sus parias*. Barcelona: Paidós.

Bauman, Zygmunt. 2017. *Modernidad líquida*. Madrid: Fondo de Cultura Económica.

Ballesteros, Jesús. 1999. «El individualismo como obstáculo a la universalidad de los derechos humanos», *Persona y derecho: Revista de fundamentación de las Instituciones Jurídicas y de Derechos Humanos* 41: 15-28.

Ballesteros, Jesús. 2003. «¿Derechos? ¿Humanos?», *Persona y derecho: Revista de fundamentación de las Instituciones Jurídicas y de Derechos Humanos* 48: 27-46.

Casilli, Antonio A. 2021. *Esperando a los robots: Investigación sobre el trabajo del clic*. Santiago de Chile: Punto de Vista.

Castellanos, Jorge. 2020. «El derecho humano a participar: estudio del artículo 21 de la Declaración Universal de Derechos Humanos», *Universitas* 30: 33-51.

Castellanos, Jorge. 2024. «Una reflexión acerca de la influencia de la inteligencia artificial en los derechos fundamentales». In *Ciencia de datos y perspectivas de la inteligencia artificial*, coordinated by Francisca Ramón, 271-300. Valencia: Tirant lo Blanch.

Fernández Ruiz-Gálvez, Encarnación. 1996. «Derechos humanos: ¿yuxtaposición o integración?», *Anuario de Filosofía del Derecho* 13: 679-702.

Fernández Ruiz-Gálvez, Encarnación. 1999. «Derechos humanos: del universalismo abstracto a la universalidad concreta», *Persona y derecho: Revista de fundamentación de las Instituciones Jurídicas y de Derechos Humanos* 41: 57-88.

Harari, Yuval Noah. 2016. *Homo deus: breve historia del mañana*. Barcelona: Debate.

MacIntyre, Alasdair. 2001. *Animales racionales y dependientes: por qué los seres humanos necesitamos las virtudes*. Barcelona: Paidós Ibérica.

Nussbaum, Martha C. 2006. *Las fronteras de la justicia: consideraciones sobre la exclusión*. Barcelona: Paidós.

Orwell, George, 2013. *1984*, Barcelona: Debolsillo.

Saiz Garitaonandia, Alberto. 2023. «Personas vulnerables, justicia e inteligencia artificial: motivos para permanecer alerta». In *Personas vulnerables y tutela penal*, directed by Norberto J. de la Mata and Ana I. Pérez Machío, 93-110. Cizur Menor: Aranzadi.

Sen, Amartya. 2000. *Desarrollo y libertad*. Barcelona: Planeta.

Terrones, Antonio L. 2023. «Inteligencia artificial fiable y vulnerabilidad: una mirada ética sobre los sesgos algorítmicos». In *Vulnerabilidad digital: desafíos y amenazas de la sociedad hiperconectada*, coordinated by Rebeca Suárez, Miguel Á. Martín and Luis M. Fernández Martínez, 263-274. Dykinson: Madrid.

Valle, Raquel. 2023. «Inteligencia artificial y derechos de las personas con discapacidad», *Revista Española de Discapacidad* 11 (1): 7-28.

Zuboff, Shoshana. 2020. *La era del capitalismo de la vigilancia: la lucha por un futuro humano frente a las nuevas fronteras del poder*. Barcelona: Paidós.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 33-50
https://doi.org/10.18543/djhr.3187 • http://djhr.revistas.deusto.es/

AI in supply chains: freedom from slavery revisited

IA en las cadenas de suministro: la libertad de la esclavitud revisada

Migle Laukyte
Universitat Pompeu Fabra. Spain
migle.laukyte@upf.edu
ORCiD: https://orcid.org/0000-0002-6331-9364

Lorena María Arismendy Mengual
CUNEF Universidad. Spain
lorena.arismendy@cunef.edu
ORCiD: https://orcid.org/0000-0002-9969-9186

https://doi.org/10.18543/djhr.3188

Submission date: 27.05.2024
Approval date: 12.09.2024
E-published: December 2024

Citation / Cómo citar: Laukyte, Migle y Arismendy, Lorena María. 2024. «AI in supply chains: freedom from slavery revisited.» *Deusto Journal of Human Rights*, n. 14: 51-71. https://doi.org/10.18543/djhr.3188

Summary: Introduction. 1. The right to freedom from slavery today. 2. Modern slavery in global supply chains. 3. AI for uncovering modern slavery and safeguarding the right to freedom from slavery. 3.1. AI and human rights. 3.2. Advancing supply chain transparency with AI. New ways to tackle modern slavery. 4. Critical aspects and final remarks. References.

Abstract: This paper addresses the link between Artificial Intelligence (AI) and the human and fundamental right to freedom from slavery: in particular, we focus on the modern slavery in global supply chains and the possibility to use AI to identify it. We analyze the slavery and its modern version, situate the AI within the human rights debate and argue that we should not only focus on how AI can violate and infringe the human rights, but also explore how AI could be useful in identifying violations and helping to combat them. We emphasize the need for inclusive datasets and stakeholder oversight and argue in support of AI to enhance transparency of international supply chains while cautioning against biases. We conclude by outlining the importance of responsible AI deployment and making a case for more regulatory efforts to protect the fundamental human right to freedom from slavery in supply chain operations.

Keywords: Artificial intelligence, Supply chains, human rights, fundamental rights, right to freedom from slavery.

Resumen: El presente trabajo trata el tema de la Inteligencia Artificial (IA) y el derecho humano y fundamental a la libertad de la esclavitud. En particular, enfocamos la esclavitud moderna en las cadenas internacionales de suministros y la posibilidad de utilizar la IA para detectarla. Analizamos la esclavitud y su versión moderna, situamos la IA dentro del debate sobre Derechos Humanos, y tratamos la idea de que ver la IA solo como una herramienta de la violación de Derechos Humanos es limitativo y que hay que explorar más como la IA podría ser útil para identificar las violaciones de los derechos humanos y para ayudarnos a combatirlas. Para lograrlo necesitamos los datos más inclusivos y la supervisión humana, y sostenemos, sin perder de vista el problema de sesgos, que la IA podría ayudar a incrementar la transparencia en las cadenas internacionales de suministros. Concluimos con la importancia del desarrollo de la IA responsable y la necesidad de más esfuerzo regulatorio para proteger este derecho en dichas cadenas.

Palabras clave: Inteligencia artificial, cadena de suministros, derechos humanos, derechos fundamentales, derecho a la libertad de la esclavitud.

Introduction

Human rights and artificial intelligence (AI) have come into clash on variety of aspects: more often than not, the AI was a tool to curtail the rights of individuals and vulnerable social groups depriving them of the little they could rely upon in terms of equal treatment, civil freedoms, social benefits, or other entitlements (among many, FRA 2022; Greiman 2021; Quintavalla and Temperman 2023).

This work addresses one of the human and fundamental rights[1] that have seldom been linked to the AI, although the situation is probably about to change. The right to freedom from slavery has been a part of our history and development as society, but was not that often related to AI. To make up for this gap and add to the existing debate, this paper focuses on modern slavery within international supply chains of everyday products, such as those of Nespresso, Starbucks, or Apple, and asks how AI can be useful to detect and combat it. We want to show that while AI poses threats to fundamental rights, it can also serve as a powerful tool for defending them. Our review of literature and analysis of the existing options for leveraging AI in the fight against modern slavery (such as AI-based due diligence monitoring using Blockchain and digital identity systems to uncover labour conditions and treatment of workers or integrating AI into automated surveillance of supply chains) supports the conclusion that a regulatory action alongside technical advancements is the right way forward. We also argue that addressing challenges of opacity and lack of transparency, human oversight, shortage of best practices and further issues are all crucial for deploying AI effectively in combating modern slavery within international supply chains.

Indeed, although the recent EU Artificial Intelligence Act (AI Act 2024)[2], the first regulation of AI in the world, addresses various issues related to fundamental rights-compliant development, deployment, and use of AI systems, including General Purpose AI (GPAI), it falls short in

[1] Although the authors are aware of the conceptual differences between human rights (Universal Declaration of Human Rights) and fundamental rights (EU Charter of Fundamental Rights), for the purposes of this paper these two terms will be used interchangeably, giving preference to the concept of fundamental rights whenever possible.

[2] We also use the definition of AI from AI Act: "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (art. 3).

fully addressing some of the fundamental rights, particularly the right to freedom from slavery. Given the EU's global influence in setting technology-related regulations (Bradford 2020), proactive measures are essential to safeguard this human right not only within the EU, but also globally. Therefore, the reference to the EU is not accidental: we believe the EU has the potential to lead the use of AI in fighting modern slavery and be an example to follow for other countries.

To properly address the topic and explain the ideas that we shortly have sketched above, the paper is organized as follows: we start with the introduction of the freedom from slavery as a human right and address the contemporary relevance of this right as a pressing human rights issue. Then we contextualize it within the contemporary production model, that is, we explain the link that exists between modern slavery and international supply chains. Then we move to explain what AI has to do with this: after briefly introducing the variety of discussions on AI and human rights, usually focusing on the threats of AI to them, we then address the possibility of using AI to prevent and identify the freedom from slavery violations that take place in the international supply chains, focusing on slavery related to work and employment conditions. Then we address critical aspects that emerged in our research and finish with conclusive remarks.

1. **The right to freedom from slavery today**

Slavery has been a part of human history probably from its beginning (Everett and Keegan 1997; Rogers 2019; Haslam 2020)[3]. If we focus on the XX century only, freedom from slavery was first recognized in The International Agreement for the Suppression of the White Slave Traffic (1904) (United Nations 2024b) and following treaties, and took a more specific form in the 1926 Slavery Convention which defined slavery as "the status or condition of a person over whom any or all of the powers attaching to the right of ownership are exercised" (United Nations 2024a). This characterization is fully supported by the Bellagio-Harvard Guidelines on the Legal Parameters of Slavery, under which the idea of control over a person by another is a key element to the definition (Allain 2012). Nonetheless, it should be

[3] One of the characters of the novel "Trust" by Hernan Diaz puts it as follows: "All throughout the history, the origin of capital was slavery. Look at this country and the modern world. Without slaves no cotton; without cotton, no industry; without industry, no finance capital. The original unnameable sin" (Diaz 2023, 299-300).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 51-71
https://doi.org/10.18543/djhr.3188 • http://djhr.revistas.deusto.es/

noted that the academic discussion on the definition of slavery is still ongoing; this persistent debate demonstrates the problems and inadequacies within the definitions provided by international treaties (Heys 2023).

The right to freedom from slavery became a part of the Universal Declaration of Human Rights (1948) and afterward was included in many other international and regional treaties. The latest of them is the Protocol to Prevent, Suppress and Punish Trafficking of Persons, Especially Women and Children (United Nations 2000). All these initiatives show that the international community took the question of slavery seriously and addressed it globally[4]. However, addressing does not mean solving. This is where this paper comes into play: is there a way to use AI for that purpose? Could AI be useful in fighting these practices of annihilation of human dignity and suppression of fundamental human freedom to make decisions on his or her life and future?

These questions will be explored in the following sections of this paper. For the time being, what matters is to highlight the need to address slavery not as a human tragedy of the past, that we have overcome and left behind. Sadly, this is not the case. Even today, certain forms of slavery persist, and shockingly, some of them are completely legal. Consider, for instance, the first section of the Thirteenth Amendment of the United States' Constitution, through which slavery can be imposed as a penalty for criminal offenses[5]. Whereas this provision is still in force at Federal level, it was only at the beginning of the XXI century that some States have banned this exception by amending their own Constitutions. Hence, it currently remains as a mostly valid legal consequence[6].

---

[4] *Cfr.* some noteworthy international provisions: Article 6.c of the Charter of the International Military Tribunal (1945), Articles 1 and 7 of the Supplementary Convention on the Abolition of Slavery the Slave Trade, and Institutions and Practices Similar to Slavery (1956), Article 8 of the International Convention on Civil and Political Rights (1966), Article 7 of the Rome Statute of the International Criminal Court (1998), Article 11 of the International Convention on the Protection of the Rights of All Migrant Workers and Members of their Families (1990), Article 3 of the Worst Forms of Child Labour Convention (1999), or Article 1 of the P029 Forced Labour Convention Protocol (2014). Some of these instruments are widely accepted, with up to 94% of countries (out of 193 countries) having ratified their content in the case of the 1999 Worst Forms of Child Labour Convention (Landman 2020, 307-310).

[5] "Neither slavery nor involuntary servitude, except as a punishment for crime whereof the party shall have been duly convicted, shall exist within the United States, or any place subject to their jurisdiction". U.S. Const. amend. XIII, § 1.

[6] It is noteworthy that among the founding NATO nations, the United States had the highest incarceration rate in 2021, with an alarming rate of 664 prisoners per

Moreover, according to the International Labor Organization (ILO) (2022), in 2021 there have been 50 million of modern slaves, that is, people who were trapped in forced labor or forced marriage.

Indeed, the concept of slavery has changed: the chains and whips –typical attributes of common imaginary of slavery of the XIX century– have been substituted by less evident but not less powerful tools of oppression of the most vulnerable ones, who, just as in the past, are exploited, sold, and treated as goods, but not as human beings.

The 1990s brought into being a new term to encompass all the novel (or not accounted for before) practices that constitute a contemporary approach towards the understanding of the phenomenon of slavery. "Modern slavery" is an umbrella term that covers the forced labor and forced marriage, child labor, domestic servitude, bond labor, organ harvesting, trafficking in persons, and sexual exploitation (Nicholson, Dang and Trodd 2018). In this paper, practices related to labor, which according to the ILO generate the largest number of (modern) slaves in the world, will be the main focus.

2. **Modern slavery in global supply chains**

This article focuses not on modern slavery as such, but on modern slavery in supply chains, that is, in this "set of upstream and downstream entities who work either directly or indirectly with the firm" (Melnyk et al. 2013). Within the international supply chains, modern slavery has multiple underlying causes. These range from poverty to racial discrimination, from corruption and criminality to inadequate laws, from unregulated business practices to societal cultural norms (Han et al. 2022, 4-5). Modern slavery has sometimes been disguised or rationalized as part of a cultural practice and life-or-death necessity of poor families who simply do not have alternatives but to rely on the work of children. For instance, India had almost 8 million children working in 2023, although it is making advancements to reduce it. This is to say, although what we qualify as modern slavery practice, in some countries is a question of survival, these practices are not acceptable for whatever reason they persist.

The supply chains that we focus on are international and involve big and powerful companies that outsource many of their functions to

100,000 population (the United Kingdom ranked second on the list, with 129 prisoners per 100,000 inhabitants). Vid. https://www.prisonpolicy.org/global/ 202.html.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 51-71
https://doi.org/10.18543/djhr.3188 • http://djhr.revistas.deusto.es/

the poor countries where the labour is cheap, and workers are vulnerable. Furthermore, labour-related rights in these contexts are also absent or unenforceable (Han et al. 2022, 7). Indeed, Gold et al. (2015, 485-494) explain that the international supply chains exploit cheap human resources, driven by global inequality and hierarchical social relations to produce goods for the global market. While cost reduction is a common goal in supply chains, in cases of slave labour, the bulk of profits are retained by "slaveholders" or businesses, and little to nothing reaches the lowest levels of the chain. The origin of slave-made commodities is concealed from the public eye, and the workers do not know or are too vulnerable to claim higher wages as well as healthy and dignified work conditions. Slave-made commodities become, therefore, mixed with other goods at subsequent supply chain tiers, such as exporters or wholesalers, before reaching consumers. These consumers are often not aware, or even prefer not to be aware, of how the products they buy have been produced. Consequently, slave labour remains largely hidden or deliberately ignored by the industrialized world.

Different abuses related to supply chain management have been brought into the light regardless the sector or industry, be it raw materials, consumer ready goods, minery, agriculture, or other sectors. For instance, and among many, the 2016 report of the International Trade Union Confederation (Howard 2016) showed the estimated hidden workforce of many of widely known companies, such as Apple, Carrefour, Nestlé, Nike, Siemens, Samsung and many others. This is to say that many of the goods and products we use every day, starting with an iPhone and finishing with a cup of hot chocolate, from the running shoes to the TV set, from food to washing machine might have been done by modern slaves.

Indeed, identifying and addressing the existence of modern slavery within the international supply chains is a complicated and challenging issue. Due to the intricate and often hidden nature of modern slavery within supply chains, it is difficult to accurately estimate the global number of individuals affected by it, as well as to appropriately tackle this rapidly expanding global problem (Meehan and Pinnington 2021, 77). Further issues ensue from the strategies to combat and stop these practices: needless to say, making big and powerful companies comply with human rights is an ongoing challenge, but not an already achieved result (United Nations 2011).

The EU is working on the legislative proposal that would prohibit the products made with forced labor on the EU market (Legislative Observatory 2022): the precedents of similar legislative initiatives have

been already included into legislative frameworks of the UK (Modern Slavery Act 2015), Australia (Modern Slavery Act 2018) and Canada (Fighting Against Forced Labour and Child Labour in Supply Chains Act 2024), among many others.

However, despite nearly a decade since the implementation of the first actual regulatory framework, the measures provided to combat modern slavery have proven to be ineffective. Consider, for instance, the UK Modern Slavery Act (2015). Section 54 of the UK Modern Slavery Act requires that companies with a business presence in the UK and an annual turnover of at least £36 million to present an annual statement on modern slavery and human trafficking (UK Government Home Office 2017). This and similar obligations are meant to outline the actions organizations have to undertake to prevent modern slavery in their business operations and international supply chains. The companies are obliged to provide details about the company's structure, business operations, and supply chains, along with policies, due diligence processes, and risk assessments related to slavery and human trafficking. They must elucidate on the effectiveness of measures taken to prevent such practices, and describe the training available to their staff, among other measures. The company's statement must be approved by the appropriate governing body and signed. Additionally, if the organization has a website, the statement must be published there, with a link on the homepage, otherwise, it must provide a copy of this statement to anyone who submits a written request. The duties outlined in the Act can be enforced by the Secretary of State through civil proceedings. Similar provisions regarding modern slavery identification and prevention can be found in the other Acts mentioned above.

Regardless of these efforts, the problem is that it remains problematic to combat modern slavery by solely promoting accountability in the public sphere through disclosure. Indeed, the UK Act –but this is not the problem of the UK Act alone– does not set forth specific reporting standards, nor does it impose penalties for non-compliance. Moreover, as long as a report is published, the company will have complied, even if it does not undertake any actions against modern slavery, therefore falling short in addressing the problem (LeBaron and Rühmkorf 2017, 20).

Furthermore, the regulation has not effectively curtailed modern slavery practices in supply chains also because influential stakeholders with vested interests often exert pressure on vulnerable workers to conceal modern slavery offenses (Yawar and Seuring 2017, 621-643). Additionally, conducting rigorous due diligence across global supply chains is a complex, time-consuming, and expensive endeavor

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 51-71
https://doi.org/10.18543/djhr.3188 • http://djhr.revistas.deusto.es/

(MacCarthy et al. 2022, 4), which the affected countries themselves nor local communities can afford. These challenges currently contribute to the persistence of modern slavery despite regulatory efforts (Mantouvalou 2018, 1017-1045; Tambe and Tambay 2020, 22).

To be sure, the companies might argue that they are not in a situation to challenge the national legal frameworks that permit these practices, and their presence offers employment where otherwise the people would starve or live even worse than they live working for the big international corporations. This is also true: yet the corporations need to act and not to close their eyes. The right way to proceed is not to continue these practices, but to work to improve the employment conditions and support the local communities by addressing specific issues and problems, such as accommodation, transportation, food, education and healthcare. As stated by Dante Pesce, the chairman of the United Nations Working Group on Business and Human Rights (Eco-Business 2018), "You [company] are, at very least, complicit if you fail to act".

3. **AI for uncovering modern slavery and safeguarding the right to freedom from slavery**

The discussion so far shows that the right of freedom from slavery remains necessary to defend; its infringements and violations are elusive, concealed, and difficult to identify. It is, therefore, hard to fight. This lack of visibility is inherent to violations of other human rights as well, such as enforced disappearances and extrajudicial killings, *e.g.*, in Mexico with over 100,000 people reported missing (United Nations 2022). Criminal groups allegedly scheme with authorities to abduct or kill individuals, while governments often fail to properly investigate cases or provide justice for victims' families. These factors ultimately allow the violations to continue unchecked and hidden. Hence, all human rights violations are difficult to investigate and bring to the light.

This is why we turn to AI: can AI be useful in this? Indeed, according to Landman (2020, 329-330), techniques, such as computational science and AI –already being used to detect and quantify several human rights violations– could be equally applicable to shedding light on modern slavery as described in the following section[7].

[7] In this regard, an initiative worth mentioning is the Human Rights Data Analysis Group (HRDAG), which has developed a Machine Learning-based tool to calculate the deaths during Syrian conflict (2011-2014), to identify where the mass graves in Mexico

But before discussing the uses of AI to identify modern slavery-aligned practices in the supply chains, we first need to situate AI within the debate on the human rights and identify what is missing in it and how addressing the modern slavery problem might push forward a change within this debate.

3.1. *AI and human rights*

The research on the impact of AI on human rights is extensive (among many, Mantelero 2022; Aizenberg and van der Hoeven 2020; European Council 2023; Jones 2023). The human rights and fundamental rights that the literature discussed and continues discussing mostly are the right to privacy and personal data protection, the right to freedom from discrimination and bias and the right to equality, the right to fair trial and other procedural rights, freedom of expression, right to healthcare and essential services, rights related to intellectual property and authorship and others.

Having said that, it is also true that not all human rights have been subject to the same attention: for example, the amount of academic literature on the threats of AI to privacy and personal data protection is many times superior to that of other rights, such as right to education, and, for the purposes of this paper, also to the right to freedom from slavery.

The question is why the freedom of slavery is not a "popular" topic for human rights scholars who work on AI. We do not know the reasons, yet we suggest that this is probably due to the lack of knowledge about AI's possibilities in this sense. Also, the lack of interest by the companies that develop AI to invest in something that might bring into the light the seriousness of the problem that they themselves have caused, could play a significant role, should their AI be based on supply chains in the first place. This is just a hypothesis that to be sure needs to be proven: however, with this work we want to contribute to bringing this possibility on the table and arguing that there should be more discussion and analysis of how to use AI in reducing and hopefully eliminating the problem of modern slavery in the world.

This work indirectly addresses one of the currently widely spread narratives on the AI. This is also proven by the aforementioned

---

are situated and to establish the patterns of discourse of human rights violators (Landman 2020).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 51-71
https://doi.org/10.18543/djhr.3188 • http://djhr.revistas.deusto.es/

literature overview: the AI is seen as a threat to human rights and as a tool to reduce humans to entities who are easy to manipulate. From this point of view, they exist to generate the data necessary to train AI systems, to buy and use services or to be processed as packages or goods in the warehouse. And this is indeed the case, as many cases of AI uses have shown. Yet the AI is not only that and should not be only that: the idea that this work defends is also related to the fact that seeing AI as a threat only is reductive. We should also start seeing AI as an opportunity: a way to use the AI in improving the situation of the most vulnerable people and to use AI for the benefit of humans and not just for the benefit of (big technological) companies.

3.2. *Advancing supply chain transparency with AI. New ways to tackle modern slavery*

To gain a deeper understanding of how AI can improve the detection of modern slavery within supply chains, it is important to examine the existing practices used for conducting them through audits.

The methods for evaluating modern slavery within international supply chains differ from conventional approaches used in regular supply chain audits. Brintrup et al. (2023, 4681) describe traditional supply chain surveillance as a "manual, and at times an opportunistic process, informed by expert knowledge and limited data. The process would involve scrutiny, validation and judgements made by a variety of supply chain professionals. For example, if a supplier's relations with competitors were of interest, the buyer might directly query the supplier or monitor industrial news sources. At other times, surveillance might be tacit. Procurement officers might collate historical data on supplier performance periodically to assist in future supplier selection. Both of these involve a degree of subjectivity and tacit human knowledge". Hence, modern slavery, as a distinct issue, requires dedicated and targeted attention to be properly addressed (Lund-Thomsen 2008, 1005; New 2015, 697; Gold et al. 2015, 10, 14).

Traditional methods for identifying these risks regularly involve customer surveys, accreditation processes, manual mapping, and monitoring of suppliers, as well as third-party auditing services (Brintrup et al. 2023, 4675). Whilst conducting any of these procedures within global international supply chains, auditors should focus on identifying specific indicators that suggest modern slavery practices might be involved. Such indicators are, for instance, the threat of physical harm

to individuals, restriction of movement, debt bondage, withholding wages, retaining passports, or the threat of denunciation to the authorities. Additional complications arise from the limited research available on this specific question and the fact that such indicators of modern slavery can be hazardous for auditors or inspectors to report, often constituting a risk to their own lives (Crane 2013, 49; Stevenson and Cole 2018, 83; Bodendorf et al. 2022, 2050-2053).

Moreover, the criminal nature of modern slavery acts committed against others, coupled with the possible severe repercussions for those involved if exposed, conveys that conventional detection and remediation practices, typically effective for other types of offenses, are often inadequate or inappropriate for addressing modern slavery. This means that, in general, regular international supply chain management practices have limited effectiveness when addressing illegal activities that are actively concealed (Gold et al. 2015, 8; Stevenson and Cole 2018, 82).

Despite the diversity in measurement strategies, Landman (2020, 330; Landman and Kersten 2016, 127) stresses that some "common themes" emerge in the process of assessing said risks, namely:

(i) All modern slavery measurement methods rely on raw data sources.
(ii) A coding or counting process transforms raw information into quantitative data, expressing different categories and dimensions of slavery.
(iii) Analytical techniques produce descriptive statistics or more complex analyses that combine or compare data across categories, variables, and dimensions.
(iv) These methods generate outputs that may improve our understanding of modern slavery, including the quantification of the total number of instances or occurrences of modern slavery risks in a given supply chain at a specific point in time (prevalence counts), explanations of such patterns, and predictions or estimates of risks of modern slavery.

Bearing the above in mind it comes as no surprise that the idea to see whether AI could offer a way forward has emerged. As of today, most of the burden of preventing, addressing, and ultimately resolving the issue of modern slavery falls on companies. Consequently, Brintrup et al. (2023, 4675) observed that several recent studies suggest that Supply Chain Digitalization (SCD) could provide companies with additional approaches to enhance existing methods for addressing visibility issues.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 51-71
https://doi.org/10.18543/djhr.3188 • http://djhr.revistas.deusto.es/

As posed by several studies, digital technologies offer substantial benefits throughout supply chain management stages in general. Theoretically, they improve demand responsiveness and capacity flexibility, helping to identify how events causing disruptions in production, sale, or distribution of products affect supply chain performance and can lead to changes in supply chain structural design and planning parameters in response to said disruptions. According to the scientific literature, technologies like big data analytics and track and trace significantly enhance data coordination and supply chain visibility for simulating and implementing recovery strategies. These findings underscore digitalization's critical role in advancing supply chain resilience and responsiveness to disruptions (Ivanov et al. 2019, 829).

Even though the use of AI is, in fact, not strictly necessary for analyzing digital data –whether derived from supply chain audits or otherwise– it could certainly offer performance enhancements compared to other methods when dealing with unstructured, large-scale digital data.

AI-powered tools are therefore being developed to aid companies in combating forced labour and other forms of modern slavery within their international supply chains. These tools leverage advanced technologies such as natural language processing, computer vision, and decision intelligence to analyze data and detect, *e.g.*, indicators of trafficking or forced labour (Li 2016, 98–99, Weinberg et al. 2020) ultimately progressing towards freedom from slavery. AI tools may also contribute to creating more transparent supply chains by analyzing large datasets from various sources to identify potential signs of modern slavery practices. The outcomes presented by this approach theoretically allow companies to take swift action to prevent and address any activity that could be related to modern slavery: that is to say, the companies have means not only to deal with concrete cases of slavery, but also identify border-line situations or conditions which could eventually lead to slavery-based relationships.

Examples of such technology involve AI-powered due diligence technology platforms like GRAT[8] or FRDM[9], which are designed to map complex supply chains and help identify and mitigate forced labour, slave labour, and human trafficking in them, possibly enabling companies to meet higher standards for anti-modern slavery legislation and human rights due diligence requirements (Nersessian and

[8] See https://counterforcedlabor.com/grat/

[9] See https://www.frdm.co/how-frdm-works

Pachamanova 2022, 2-46). In addition to that, these platforms also educate, create more awareness and also help the companies to track their advancements in dealing with these problems: besides identifying risks of forced labour, these systems also track progress of how these risks are being dealt with and provide with support and assistance in finding the most appropriate solution.

Indeed, the adoption of AI tools has emerged as a way to combat modern slavery: should companies choose to explore it, they have a possibility to boost their existing slavery prohibition due diligence processes using this technology, for instance, along with blockchain technology to record modern slavery in the supply chain offenses publicly and transparently (Tambe and Tambay 2020). As a result, prompt action could be taken based on the accurate outputs generated by the enhanced auditing process. Despite the importance of explainability in algorithmic results –subsequently addressed in this section– practitioners often prioritize accuracy over explainability. There are, however, significant differences in this preference across various industrial sectors and application areas (Brintrup et al. 2023, 4674).

After a thorough literature review on this issue, Han et el. (2022, 16) found that scientific studies conducted on the use of digital technologies –such as AI, cloud computing, and biometric identification– specifically aiming to reduce modern slavery risks within supply chains, currently do not address how these technologies achieve this objective. This suggests that their ability to boost supply chain transparency is not necessarily well-understood, nor has it been unequivocally demonstrated. Nevertheless, in the context of Industry 4.0, these advanced digital tools hold promise as potential solutions for combating modern slavery, even though further research is needed to assess their effectiveness and associated challenges and eventually identify where these technologies fall short to achieve their objectives.

As a part of the challenges mentioned above, it is crucial to recognize risks associated with supply chain digitalization and AI powered surveillance in general, and with modern slavery practices in particular. Among many, a rising concern regarding AI systems' ability to comply with individuals' privacy are among the most prominent issues. It should be noted that EU companies have to comply with 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 (GDPR). For example, in accordance with article 13 GDPR, where personal data relating to a data subject are collected from the data subject, auditors or companies,

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 51-71
https://doi.org/10.18543/djhr.3188 • http://djhr.revistas.deusto.es/

using AI tools, should inform the data subject of the existence of automated decision-making, as well as meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. This conflicts directly with the so-called *black box* effect of AI algorithms.

To be sure, privacy and personal data protection legislation differs and there is no harmonized approach to it outside the EU: yet, as already mentioned before, in this case, as well as in other cases, the EU's approach or the so called "Brussels effect" (Bradford 2020) could be a starting point in terms of understanding and treating personal information of the workers.

Further issue is related to biases: regarding the data that the algorithm used to detect modern slavery in supply chains is based on, there could be several biases influencing its decision-making process. Said biases may arise from several sources, namely: (i) data imbalance, such as overrepresentation, or underrepresentation of certain phenomena in the datasets; (ii) incorrect input data, such as extracting data from sources that may not be sufficiently reliable like *mock reports* where suppliers only appear to be doing the right things on audit day; (iii) irrelevant reports where companies disclose certain types of unrelated information, including supply chain membership, labour policies, environmental impact, and social information; (iv) social media posts, or the news (Stevenson and Cole 2018, 85).

Furthermore, many AI algorithms make use of "black box" methods, which pose interpretation difficulties, making it problematical to understand the reasoning behind specific outcomes or predictions (among many, Pascuale 2025; O'Neil 2017; Koivisto 2021; Brkan and Bonnet 2019; Veale and Brass 2019). These could lead AI systems to analyse data preferring some outcomes over others. For example, if the data is skewed towards certain regions or sectors, the model may underestimate risks in other areas; an algorithm may associate certain worker demographics with higher modern slavery risks, even if that correlation is driven by systemic inequalities rather than actual risk factors. All of these factors certainly highlight the need for human oversight to help mitigate risks (Brintrup et al. 2023, 4681-4685). Yet, even the most attentive and professional human oversight clashes with the aforementioned lack of transparency and explainability, two factors that complicate the process of making any corrections to possible algorithmic errors or biases in the modern slavery risk assessment. In addition to that, other factors, such as publicity, transparency, traceability, explainability, and auditability of these algorithms are essential to prevent the violation of fundamental rights, including

equality, privacy or the right to non-discrimination (Iturmendi 2023, 266-269).

4. **Critical aspects and final remarks**

In this article we have briefly described the phenomenon of the modern slavery and situated it within the international supply chains of multinational businesses. We have also described the importance to change the perspective on AI: it can be used to identify these illegal practices. We have also warned that while AI holds promise in enhancing transparency and risk assessment in supply chains, it also carries risks of bias and perpetuating existing inequalities if not designed and deployed carefully.

To be sure, we are not suggesting that the AI alone will make the slavery disappear. Far from that: what we claim is that there are ways to make the AI useful to help us fighting against those who submit human beings to unbearable torture and depravation of human dignity. We argue that the AI is part of a solution, but not a solution in itself.

Furthermore, we are also aware that what AI can do is to identify the illegal practices, but cannot actually stop them: further steps have to be taken by those in charge as knowledge slavery-friendly practices alone do not affect these practices. Knowledge is just the first step and further procedures have to be created to effectively stop them and prevent them from being re-established.

As identified in this paper, there is an additional problem related to the source of AI: as usual, AI is generated mainly by the private sector. We should foster and support the development of such systems by public or non-profit sectors. Indeed, if we want to identify those companies that do not comply with the regulations, that knowingly and consciously have built their international supply chains on abuses and that generate profits for the detriment of human rights and freedoms, we should rely on free and independent (open source) AI systems and platforms.

This need for independent technologies fits within a wider framework of needs related to stability and resilience of democratic states vis-a-vis large technological companies. Upholding and safeguarding fundamental and human rights, including the right to freedom from slavery, is a core responsibility of governments. Private sector involvement in combating modern slavery could be influenced by corporate interests, market dynamics, and financial motivations. This might lead to selective prioritization of certain aspects of anti-slavery

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 51-71
https://doi.org/10.18543/djhr.3188 • http://djhr.revistas.deusto.es/

efforts that align with business goals, potentially overlooking broader systemic issues or marginalized groups. When this responsibility is largely delegated to private entities, it can shift accountability away from democratic institutions that are designed to represent and protect the interests of citizens. If the fight against modern slavery becomes privatized, human rights, including freedom from slavery, may be treated as commodities, subject to market forces rather than inherent rights that every individual possesses as a human being.

On a tangentially related note, most AI advancements aimed at promoting freedom from slavery are designed to ensure compliance with current regulatory standards. Consequently, if legal frameworks evolve to better safeguard human rights, innovation will likely align with these updated standards. The focus on compliance with current regulatory standards reflects the practical approach of AI developers and organizations working in the anti-modern slavery area. By prioritizing adherence to existing legal frameworks, AI technologies are only designed to address immediate needs and meet specific requirements related to detecting and preventing modern slavery. This is not to imply that current developments addressed in this study are inherently negative, but rather to emphasize the need for additional actions and interventions to address modern slavery and protect vulnerable individuals.

Another essential aspect to keep in mind in using AI for international supply chain control, is that the mechanisms for accountability and oversight are critical. The reviewed literature suggests that transparency in AI systems operations and reliability on the results are crucial, if we want to shift the perception of AI from a threat to an opportunity.

To be sure, the use of AI in fighting the modern slavery should not compromise other human rights, such as right to privacy or protection from personal data abuse. Nor should it turn into unregulated mass surveillance and data harvesting from developing countries. Ethical frameworks should be discussed and aligned between different countries and cultures, prioritizing the most beneficial choices for the most vulnerable when balancing these values.

As a mere (yet powerful!) tool in this endeavour, AI can help aggregate and analyse data to support modern slavery risk assessment. Though literature on this topic is scarce, the most relevant studies agree on this. However, it requires accompaniment, monitoring and joint management overseen by diverse stakeholders, such as businesses, governments, civil society organizations, and worker representatives, with governments assuming a bigger role and leading

this fight. This oversight ensures inclusive and representative datasets and reliable outcomes of AI systems, which can help reduce modern slavery.

References

AI Act. 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))*. Accessed May 24, 2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html

Allain, Jean, ed. 2012. *The legal understanding of slavery: From the historical to the contemporary*. Oxford: Oxford University Press.

Aizenberg, Evgeni and van der Hoeven, Jeroen. 2020. «Designing for the Human Rights in AI», *Big Data and Society*, 7(2). Accessed May 13, 2024. https://journals.sagepub.com/doi/10.1177/ 2053951720949566

Bodendorf, Frank, Fabian Wonn, Kristin Simon and Jörg Franke. 2022. «Indicators and countermeasures of modern slavery in global supply chains: Pathway to a social supply chain management framework» *Business Strategy and the Environment* 32 (4): 2049–2077. Doi: 10.1002/bse.3236.

Bradford, Anu. 2020. *The Brussels effect: How the European Union rules the world*. Oxford: Oxford University Press.

Brintrup, Alexandra, Edward Kosasih, Philipp Schaffer, Ge Zheng, Guven Demirel y Bart L. MacCarthy. 2023. «Digital supply chain surveillance using artificial intelligence: definitions, opportunities and risks», *International Journal of Production Research* 62 (13): 4674–4695. Doi: 10.1080/00207543.2023.2270719

Brkan Maja and Grégory Bonnet. 2019. «Legal and technical feasibility of the GDPR's quest for explanation of algorithmic decisions: of black boxes, white boxes and fata morganas», *European Journal of Risk Regulation* 11: 18-50.

Crane, Andrew. 2013. «Modern slavery as a management practice: exploring the conditions and capabilities for human exploitation», *Academy of Management Review* 38 (1): 49-69.

Diaz, Hernan. 2023. *Trust*. London: Pan Macmillan.

Eco-Business. 2018. «Work with suppliers, not against them, to end modern slavery.» *Eco-Business*, November 14. Accessed May 27, 2024. https://www.eco-business.com/news/work-with-suppliers-not-against-them-to-end-modern-slavery/

European Council. 2023. «Human rights by design - future-proofing human rights protection in the era of AI». Accessed May 23, 2024. https://www.coe.int/en/web/commissioner/thematic-work/digital-technologies

Everett, Susanne and Susanne Keegan. 1997. *History of Slavery*. Rochester: Grange Books.

Fighting Against Forced Labour and Child Labour in Supply Chains Act 2024. Accessed September 16, 2024. https://laws.justice.gc.ca/eng/acts/F-10.6/

Fundamental Rights Agency (FRA). 2022. *Bias in algorithms - Artificial intelligence and discrimination*, Accessed May 23, 2024. https://fra.europa.eu/en/publication/2022/bias-algorithm

Gold, Stefan, Alexander Trautrims, and Zoe Trodd. 2015. «Modern slavery challenges to supply chain management», *Supply Chain Management* 20 (5): 485-494. Doi: 10.1108/SCM-02-2015-0046

Greiman, Virginia. 2021. «Human rights and artificial intelligence: a universal challenge», *Journal of Information Warfare* 20 (1): 50-62.

Han, Chen, Fu Jia, Menggi Jiang, and Lujie Chen. 2022. «Modern slavery in supply chains: a systematic literature review», *International Journal of Logistics Research and Applications* 27 (7): 1206–27. Doi: 10.1080/13675567.2022.2118696.

Haslam, Emily. 2020. *The slave trade, abolition and the long history of international criminal law*. Oxon: Routledge.

Heys, Alicia. 2023. *From conflict to modern slavery: The drivers and the deterrents*. Oxford: Oxford University Press.

Howard, Neil. 2016. «Scandal: Inside the global supply chains of 50 top companies». *International Trade Union Confederation*. Accessed August 25, 2024. https://www.ituc-csi.org/IMG/pdf/pdffrontlines_scandal_en-2.pdf

International Labor Organization. 2022. *Walk free, and International Organization for Migration (IOM), Report on Global Estimates of Modern Slavery. Forced Labour and Forced Marriage, Geneva, September of 2022*. Accessed April 22, 2024. https://www.ilo.org/wcmsp5/groups/public/---ed_norm/-- ipec/documents/publication/wcms_854733.pdf

Iturmendi, José M. 2023. «La discriminación algorítmica y su impacto en la dignidad de la persona y los derechos humanos. Especial referencia a los inmigrantes.» *Deusto Journal of Human Rights* 12: 257-284.

Ivanov, Dimitry, Alexandre Dolgui, and Boris Sokolov. 2019. «The impact of digital technology and industry 4.0 on the ripple effect and supply chain risk analytics», *International Journal of Production Research* 57 (3): 829-846.

Jones, Kate. 2023. «AI governance and human rights. Resetting the relationship», *Research Paper of Chatham House.* Accessed May 12, 2024. https://www.chathamhouse.org/sites/default/files/2023-01/2023-01-10-AI-governance-human-rights-jones.pdf

Koivisto Ida. 2021. «The digital rear window: Epistemologies of digital transparency», *Critical Analysis of Law* 8 (1): 64-80.

Landman, Todd. 2020. «Measuring modern slavery: Law, human rights, and new forms of data», *Human Rights Quarterly* 42 (2): 303-331.

Landman, Todd and Larissa Kersten. 2016. «Measuring and monitoring human rights», in *Human Rights: Politics and Practice,* edited by Michael Goodhart, 127-144. Oxford: Oxford University Press.

LeBaron, Genevieve and Andreas Rühmkorf. 2017. «Steering CSR through home state regulation: A comparison of the impact of the UK bribery act and modern slavery act on global supply chain governance», *Global Policy* 8: 15-28.

Legislative Observatory. 2022. 2022/0269(COD). *Prohibiting products made with forced labour on the Union market*. Accessed August 29, 2024. https://oeil.secure.europarl.europa.eu/oeil/ popups/ficheprocedure.do?lang=en&reference=2022/0269(COD).

Li, Peng Lai. 2016. «Natural language processing», *Georgetown Law Technology Review* 1 (1): 98-104.

Lund-Thomsen, Peter. 2008. «The global sourcing and codes of conduct debate: five myths and five recommendations», *Development and Change* 39 (6): 1005-1018.

MacCarthy, Bart, Ahmed Wafaa, and Demirel Guven. 2022. «Mapping the supply chain: why, what and how?», *International Journal of Production Economics* 250 (108688): 1-20. Doi: 10.1016/j.ijpe.2022.108688

Mantelero, Alessandro. 2022. *Beyond data: Human rights, ethical and social impact Assessment of AI*. Amsterdam: Asser Press.

Mantouvalou, Virginia. 2018. «The UK modern slavery act 2015 three years on», *The Modern Law Review* 81 (6): 1017–1045.

Meehan, Joanne, and Bruce D. Pinnington. 2021. «Modern Slavery in Supply Chains: Insights Through Strategic Ambiguity», *International Journal of Operations & Production Management* 41 (2): 77-101.

Melnyk, Steven A., Ram Narasimhan, and Hugo DeCampos. 2013. «Supply chain design: Issues, challenges, frameworks and solutions». *International Journal of Production Research* 52 (7): 1887–1896.

Modern Slavery Act 2015. Accessed September 16, 2024. https://www.legislation.gov.uk/ukpga/2015/30/contents

Modern Slavery Act 2018. Accessed September 16, 2024. https://www.legislation.gov.au/C2018A00153/latest/text

Nersessian, David and Dessislava Pachamanova. 2022. «Human trafficking in the global supply chain: Using machine learning to enhance understanding of corporate disclosures under the UK modern slavery act», *Harvard Human Rights Journal* 35: 1-46.

New, Steve. 2015. «Modern slavery and the supply chain: the limits of corporate social responsibility?», *Supply Chain Management: An International Journal* 20 (6): 697-707.

Nicholson, Andrea, Minh Dang, and Zoe Trodd. 2018. «A full freedom: Contemporary survivors' definitions of slavery», *Human Rights Law Review* 18 (4): 689–704.

O'Neil, Cathy. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. London: Penguin Random House.

Pasquale, Frank. 2015. *The black box society*. London: Harvard University Press.

Quintavalla, Alberto and Jeroen Temperman, eds. 2023. *Artificial intelligence and human rights*. Oxford: Oxford University Press.

Rogers, Glenn. 2019. *A Brief History of World Slavery*. Abilene, TX: Simpson & Brook Publishers.

Stevenson, Mark and Rosanna Cole. 2018, «Modern slavery in supply chains: a secondary data analysis of detection, remediation and disclosure», *Supply Chain Management* 23(2): 81-99.

Tambe, Pratap and Prerna Tambay. 2020. «Reducing modern slavery using AI and blockchain», *IEEE / ITU International conference on artificial intelligence for good (AI4G)*: 22-27.

UK Government Home Office. 2017. *Transparency in supply chains, etc. A practical guide*. Accessed September 16, 2024. https://assets.publishing.service.gov.uk/media/61b7401d8fa8f5037778c389/Transparency_in_Supply_Chains_A_Practical_Guide_2017_final.pdf

United Nations. 2024a. *Slavery Convention. Adopted September 25, 1926.* Accessed May 17, 2024. https://www.ohchr.org/en/instruments-mechanisms/ instruments/slavery-convention.

United Nations. 2024b. *The International Agreement for the Suppression of the "White Slave Traffic"*. Paris, 18 May 1904. Accessed: May 18, 2024. https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=VII-8&chapter=7&clang=_en.

United Nations. 2022. *Mexico: Dark landmark of 100,000 disappearances reflects pattern of impunity, UN experts warn*. Statement. Accessed August 29, 2024. https://www.ohchr.org/en/ statements/2022/05/mexico-dark-landmark-100000-disappearances-reflects-pattern-impunity-un-experts

United Nations. 2011. *Guiding principles on business and human rights*. Accessed May 13, 2024 https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

United Nations. 2000. *Protocol to prevent, suppress and punish trafficking of persons, especially women and children*. Accessed August 23, 2024. https://www.ohchr.org/en/instruments-mechanisms/instruments/protocol-prevent-suppress-and-punish-trafficking-persons

Veale Michael and Irina Brass. 2019. «Administration by algorithm? Public management meets public sector machine learning» In *Algorithmic Regulation,* edited by Karen Yeung and Martin Lodge, 121-149. Oxford: Oxford University Press.

Weinberg, Nyasha *et al*, 2020. «AI against modern slavery: Digital insights into modern slavery reporting. Challenges and opportunities», *Proceedings of the AAAI Fall Symposium on AI for Social Good Virtual Symposium*, *CEUR Workshop Proceedings,* 2884: 1-8. https://cdn.walkfree.org/content/uploads/2021/05/05173207/20210428-digital-insights-into-modern-slavery.pdf

Yawar, Sadaat Ali, and Stefan Seuring. 2017. «Management of social issues in supply chains: a literature review exploring social issues, actions and performance outcomes», *Journal of Business Ethics* 141 (3): 621-643.

The systematics of the European Artificial Intelligence Act in the context of the fundamental rights of the Union: the myth of the digital constitutionalism

La sistemática del Reglamento Europeo de Inteligencia artificial en el contexto de los derechos fundamentales de la Unión: el mito del constitucionalismo digital

Ainhoa Lasa López
University of the Basque Country, Spain
ainhoa.lasa@ehu.eus
ORCiD: https://orcid.org/0000-0003-1417-0185

https://doi.org/10.18543/djhr.3189

Submission date: 28.05.2024
Approval date: 12.09.2024
E-published: December 2024

Citation / Cómo citar: Lasa, Ainhoa. 2024. «The systematics of the European Artificial Intelligence Act in the context of the fundamental rights of the Union: the myth of the digital constitutionalism.» *Deusto Journal of Human Rights*, n. 14: 73-100. https://doi.org/10.18543/djhr.3189

Summary: Introduction: AI as a form of production of semiotic capitalism. 1. Some considerations on digital materiality and its legal-political implications. 2. Constitutional paradigms and AI. 2.1. Digitalisation: the liquefaction of analogue legal orders? Or a new constitutional paradigm? 2.2. Analysis of the AI Act from the material-constitutional system of EU Law. 2.3. The antijuricity(?) of the «digital» subject. Final conclusions: the «constitutional» inconsistencies of digital constitutionalism. References.

Abstract: In recent years, Artificial Intelligence (AI) bases on data driven and machine learning have been at the centre of debates on the implications of certain uses of this technology on fundamental rights in terms of individual and social risks. At the national level, reflections on whether or not AI systems have their own ontological determinism seem to have come up against the obstacles of the staticity of constitutional frameworks that are still analogical. In the European legal order, the most disruptive digital effects of the so-called knowledge economy on the subject and his or her rights seem to be conditioned by the telos of the centrality of the human being in his/her objective-axial dimension (guarantee of the Union's values) and subjective dimension (protection of the Union's fundamental rights). The European Union

Artificial Intelligence Act would be its most recent legal-normative concretisation, in line with other norms of secondary law that would outline the dynamics of the so-called digital constitutionalism.

Key words: homo digitalis, historicity of the subject, digital rights, technological power, form of production, form of existence, normative irrationality.

Resumen: La Inteligencia Artificial (IA) basada en datos y el aprendizaje automático ha centralizado en los últimos años los distintos debates sobre las implicaciones de determinados usos de esta tecnología en los derechos fundamentales en términos de riesgos individuales y sociales. En el plano nacional, las reflexiones en torno a la posesión o no de un determinismo propio ontológico de los sistemas de IA parecen haberse topado con los obstáculos de la estaticidad de unos marcos constitucionales todavía analógicos. En el orden jurídico europeo, los efectos digitales más disruptivos de la denominada economía del conocimiento para el sujeto y sus derechos parecen condicionarse por el telos de la centralidad del ser humano en su dimensión objetivo-axial (garantía de los valores de la Unión) y subjetiva (tutela de los derechos fundamentales de la Unión). El Reglamento Europeo de IA sería su concreción jurídico-normativa más reciente, cohonestándose con otras normas de derecho secundario que trazarían las dinámicas del denominado constitucionalismo digital.

Palabras claves: homo digitalis, historicidad del sujeto, derechos digitales, poder tecnológico, forma de producción, forma de existencia, irracionalidad normativa.

Introduction: AI as a form of production of semiotic capitalism[1]

The aim of this paper is to analyse whether the European Union (EU) Artificial Intelligence Act (AI Act)[2] can be configured as a paradigm of a rights-based model of regulatory production on digitalisation, following Bradford's classification[3] (2023, 105-145); or, on the contrary, whether it is limited to regulating a new form of commodity production (in the form of AI systems[4]) whose potential impacts on EU fundamental rights and values are modulated by the structural and structuralising guarantee of the materiality of the European legal order: internal market, free competition and fundamental economic freedoms.

The difference is not superficial, while in the first interpretation rights and the axial are configured as insurmountable barriers to the

---

[1] This work is an extended version of the communication presented at the ICON•S (International Society of Public Law) Annual Conference 2024, «The Future of Public Law: Resilience, Sustainability, and Artificial Intelligence», held in Madrid on 8-10 July. Moreover, it has been carried out in the framework of activities of the following research project: "Biosurveillance through Artificial Intelligence (AI) in the post COVID era: Corporality, Identity and Fundamental Rights" (TED Code 2021-129975B-C21), Main researcher: Leire Escajedo San Epifanio.

[2] On 21 May 2024, the Council of the European Union approved the compromise amendments of the IA Act (https://www.consilium.europa.eu/es/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/), which had already been endorsed by the European Parliament on 24 March of the same year (https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf). This concluded the trilogue process started three years earlier, in 2021, when the European Commission presented the legislative proposal (COM (2021) 206 final), https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

[3] The author approaches the unstoppable advance of the degrees of technological realisation from the consideration of digitalisation as a social fact of geopolitical scope, hence the recourse to the terminology of "digital empires", thus illustrating, therefore, the institutionalised technological vision as an element that can go in different directions according to the objectives of the political actors who embody its development trajectories. To this end, she traces the differences among the market-driven US regulatory model, the Chinese model of state regulation and the already referenced normative-"iusfundamental" model of the European Union. While acknowledging that the latter does not entirely escape indirect regulation or politically imposed deregulation of the market. Consequently, at the present stage the difference between Western and Eurasian realities seems to be that the former adopts the superstructures of pluralism, liberalism and individualism to clothe the substance of technocracy, while the latter is guided by a political model that tends to be autocratic/liberal or communitarian.

[4] The AI Act establishes a regulation of AI systems and models based on risk (unacceptable, high, low or non-existent), being classified according to the degree of risk and its impact into: prohibited AI systems (Article 5), high-risk AI systems (Article 6), general purpose AI models with systemic risk (Article 51).

instrumental and calculating techno-scientific rationality of AI that prevent the subject and his or her rights from being shaped by the ontology of human capital; in the second approach, from the intertwining of the production form of AI, as a commodity form, with the legal form of the Act whose iusfundamental and axial bases are accessory to the legal basis of the market, the individual becomes a mere incarnation of an abstract and impersonal subject of rights (user/consumer), a pure product of the social relations of digitalisation as a form of existence.

The interpretation of AI as a form of existence implies attributing to it the identity of a project of anthropological change, in the sense of a new existential identity. The transition from the analogical human to the digital human or *homo digitalis* (Han 2014, 28) that possesses an immediate and automated knowledge of reality, increasing his/her relational intelligence. The relationship as a meta-category, the absolute ontological principle that theorises the generalised connection of all living beings, of all material and immaterial reality, in a communication/information network that grants identity to technological singularity (Cristianini 2023).

This characterisation as a form of existence implies situating AI within the framework of the new knowledge economy, where data science and the development of AI lead to theorising that the world is a massive information process. So we are mutually connected informational organisms and part of an informational environment (the infosphere) that we share with other informational agents, natural and artificial, which process information logically and autonomously (Floridi 2017, 79-106). This claimed independence and dissociation of the mind from the body, that a thought, to be such, does not need the processing of sensory data generating the exchange between information and knowledge, transforms the former, information, into chains of signs that are processed obeying the primary principle of non-contradiction and the formal rules of mathematical calculus (Finelli 2022).

Information thus distorted (independence and dissociation of the mind from the body, or disconnection of verbal language from natural language, a thought that, abstracted from the body, can only compose abstract and universal codes) generates a knowledge composed of connections and calculation operations that formally avoids discursive contradiction. A digitalised knowledge that is nothing but the product/commodity of digital capitalism or semiotic capitalism, insofar as infinite flows of signifiers, of disembodied signs, break all links with the real referent, mutually exchanging each other (Berardi 2021, 31).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

1. Some considerations on digital materiality and its legal-political implications

The main characteristic of AI and the computational logics that support it is that of formalism: that is, of a syntax that, by means of mathematical and topological rules, writes and rewrites, according to the levels of processing and calculation, a set of signs (Finelli 2022). However, the chains, topologies and spatial architectures of signs that function according to a syntax of precise rules of displacement and calculation do not themselves possess a semantics, which must also be assigned to them by the specific intentions and utilities of the programmer who establishes and organises the modality of a given accumulation and processing of data (Numerico 2021, 25-55). Hence, computerised information is not autonomous. It needs to be interpreted, endowed with a meaning that does not derive from the formalism of computational rules. In other words, algorithms are automatic processes, but programmed by human beings, with mathematical programming or data selection being the materiality where the semantics of meanings are used[5].

A semantics of meanings that leads us, in turn, to rethink two intertwined dynamics in the framework of AI: on the one hand, what I call the externality of AI, which leads to the extractivist practices of data on which weak (deduction) or deep (reasoning-planning-compension of causality) learning is fed (Mitchell 2024, 32:57); on the other hand, the internality of AI, understood as a process of extracting information value from data through algorithms that will reproduce and amplify the analogical inequalities present in legal and social orders (Crawford 2023, 195). Both dynamics belong to the logic of materiality, they are tangible because they are framed in the strategy of capital accumulation, however much they pretend to be represented as abstract moments disconnected from their concrete projections, for they are still mechanisms of valorisation, forms of power and politics that they present as legal measurements/rules of objective relations.

For this reason, I do not share those theses that refute the implications of AI in the configuration of the legal world, redirecting its

[5] "In the communicative spaces designed by the big technology companies and ordered through algorithms, the digital culture that is being imposed is based on uncertainty… The confusion that these algorithms are causing in communicative processes and in the public space has no known historical precedent. The result is the weakening, if not the destruction, of the shared social perception of reality that existed before the digital society" (Balaguer and Balaguer 2023, 78).

field of action only to the spaces of politics, as if the legal were nothing more than the concretisation of legislative political development. A 'Political Constitution' redirected to the space of decisional politics and a 'Legal Constitution' whose formal guarantees are fortresses impenetrable by the technological context. On the contrary, I consider AI to be the result of a precise development of a new mode of production of the social order of capital in its digitalised phase, which delegates to technology the new form of production/extraction of value, of accumulative profit, automating human lives and transferring the new form of production to the communicative processes themselves. Messages, data, as commodities are the product of the productive mechanisms of AI that 'discovers' correlations and extract rules. AI is therefore a creator of rules that construct a representation with not only political, but also juridical effects (Chiaritti 2021, 15). There is no political-economy or state-society split that makes it possible to immunise the legal from digital reality.

Neutrality towards the digital in contemporary constitutional states is only possible if the digital is relativised as a *locus naturalis*, as it was during the liberal state with the market. In a different sense, if we analyse the structural isomorphism of capital and the digital (in which AI is situated) from the interrelations between Power-Law and forms of states, we come to the conclusion that there is no spontaneous emergence of the digital fact as a natural fact detached from its relevant factual-normative context, since it lacks autonomous consistency in itself.

From this perspective that shapes AI in the digitalisation as a process of relational social action of capitalism that endlessly accumulates electronic and digital transactions generating a spatio-temporal rupture, I intend to emphasise the concomitance of digitalised capitalism with the unlimited expansion of the global power of the market. The objectives of efficiency, of amplifying human rationality in decision-making through the use of algorithms, are instrumental to the economic objective of profit maximisation. AI is thus embedded in an interpretation of digitalisation as another process of the advance of the market in its financialised form that began to take shape in the late 1970s, when the expectation of unlimited growth in the rate of profit began to stagnate (Betancourt 2015, 215-224).

Digital capitalism and its legal-political connotations do not, as I have pointed out, reactivate a state-society separation, like the liberal state. In any case, we are witnessing a recomposition of political-economic relations that produces a reconversion of society into a 'digital performance society', as the source or raw material of digitised

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

power legally mediated by Law in the form of the global power of the market. This connection between the factual and the legal reproduces the legal mediation of social relations in the different forms of state or institutionalised relational compositions between the state and the market, such as the current market constitutionalism, where statehood as a symbiotic framework of global market power accentuates the juridification of the spaces of control of the conflicts that hinder the centrality of the market through technological and financial progress. In accordance with this last premise, I contextualise the so-called technological challenges to analogical constitutional spaces and their reconduction to the supranational space of the Union in order to address the moment of control of system providers and those responsible for the deployment of AI systems, which has been substantiated in the AI Act. Interrelational ensemble, global market power in its digital form, market Law - market-state form, which will be dealt with in the second section to elucidate the legal nature and constitutional (digital) scope of the AI Act.

In parallel, the recitals of this EU secondary legislation emphasise the political and legal determinism of AI as a human-centred technology, giving it an anthropocentric telos where AI systems are tools whose ultimate goal is to enhance human welfare. However, algorithms merely analyse the relationships in the data, not the values or meaning they represent. In other words, they act the relational form by questioning, only in appearance, relational materiality because they are generators of their own matter or substance. Concretely, digital space is the digitalised form of the once physical factories, the space of organisation and management of digital capital. The producers of algorithmic subjectivisation are the digital users in the form of the transfer of large masses of (economically relevant) information that are captured on the global platforms of Google, Apple, Facebook, Amazon and Microsoft, vastly increasing their market shares (and profits) in key segments such as advertising and data retailing to third parties. In this space we no longer capitalise on things, objects, products, coins, banknotes, but on personal information regarding our feelings, our desires, our emotions, our behaviour (Zuboff 2020, 315-364).

The process of digitalisation that has characterised recent decades corresponds to a new phase of capital and capitalism that is no longer based on the accumulation of money, whether real or virtual, but on big data. In short, when we do a Google search, look at Wikipedia or make a purchase on Amazon, we move an infinite amount of data that in its sheer movement emulates that of capital and is self-valorising. Capital and the digital are operators of interactions between

individuals and, therefore, imply a relationship that develops its own subjectivity[6].

This digital social order generates its own subjectivity because the digital connection is an economic activity carried out by a multiplicity of subjects who are situated within the framework of an economic activity, acquiring the status of economic subjects (suppliers, those responsible for deployment, importers and distributors, users, consumers) under the guarantee of economic freedoms that in EU Law are configured as fundamental rights. These economic freedoms, and not the fundamental rights of the Constitutional State, are those that shape the parameters of development of the subjects in their digital interactions and their concretisations, in terms of privacy and property. Economic freedoms as negative freedoms are the ones that determine the spaces and limits for the exercise and effectiveness of rights. Positive freedom, the equality of positions, becomes unrealisable in the constitutional paradigm of the market in its digital form. To the study of these questions and their projection in the EU's fundamental values and rights, to which AI Act is circumscribed, we will also dedicate the second of the sections, questioning the legality of the digital subject in the constitutional state of law.

On the other hand, by reasoning about AI and its mode of subjective production, it seems that I intend to reproduce the transition from ownership to appropriation of an immaterial good (data) by digital capitalism. However, what is relevant is not the reproduction of the digital social order, understood as the new modes of production of 'profits', in the sense of recital 4 of the AI Act[7]; but the moment of power as control[8] over who decides and on the basis of what they

---

[6] "Our basic illusion is that big data appears to us as a substance, a kind of magical natural resource to be extracted from a mine: we even use terms like data mining to consolidate this fantasy of ours, mimicking the same mechanisms that underpinned the so-called first industrial revolution" (Lanier 2013, 131).

[7] "AI is a fast-evolving family of technologies that contributes to a wide array of economic, environmental and societal benefits across the entire spectrum of industries and social activities. By improving prediction, optimising operations and resource allocation, and personalising digital solutions available for individuals and organisations, the use of AI can provide key competitive advantages to undertakings and support socially and environmentally beneficial outcomes, for example in healthcare, agriculture, food safety, education and training, media, sports, culture, infrastructure management, energy, transport and logistics, public services, security, justice, resource and energy efficiency, environmental monitoring, the conservation and restoration of biodiversity and ecosystems and climate change mitigation and adaptation".

[8] "Because the truth is that many of the decisions that are adopted in the spaces of uncertainty opened up by technological development have an ostensible legal

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

decide, because control, if it is emptied of conflict, of the social question, is emptied in the specific form of a technocratic governance or technocracy of numbers. To avoid this, they appeal to the guarantees of a digital juridification or mediation of digital singularity through regulation by design (obligation to programme or codify the technology so that it complies with certain legal obligations) in the form of co-regulation (Van Cleynenbreugel 2022, 203-205). A regulatory framework involving both digital operators and public institutions (the European AI Office under the Commission, the European AI Board, scientific panel of independent experts, national competent authorities… Chapter VII. Governance of the AI Act) in the establishment, implementation or enforcement of regulatory standards with the objective of achieving "a uniform legal framework, in particular for the development, the placing on the market, the putting into service and the use of AI systems in the Union" (Recital 1 of the AI Act). However, this technical harmonisation, which is brought back to the question of power, control and law in the European legal order, is traced within the contours of a relationship between politics and economics juridified by a constitutional paradigm, that of the market, which is confronted with the fideisms of a constitutionalism characterised as digital under the institutionalised form of the constitutional state. The last of the sections of this contribution is devoted to exploring this confrontation.

2. **Constitutional paradigms and AI**

Data-driven AI and machine learning have in recent years been at the centre of various constitutional debates on the implications of certain uses of this technology on fundamental rights in terms of individual and societal risks (Simoncini and Longo 2022, 27-41). At the state level or at the level of domestic legal systems, the reflections have essentially focused on the effects of AI systems on the rights of people in their individual and collective dimensions, warning of the difficulties of addressing the digital challenges for the subject and his/her rights from the frameworks of still analogical constitutions (Presno 2022). From these limitations, methodologies of analysis have been articulated which, in their most finished formulations, redirect the possible

significance and conflict as soon as they are situated on the line where diverse rights, values and interests converge, worthy of legal protection and often in conflict" (Esteve and Tejada 2013, 30).

solutions under the legal form of the 'Constitution of the algorithm', harmonising the regulatory algorithms of the digital reality with the constitutional principles and values, in order to modernise, digitalising, the constitutional device to the new conditions of the digital era (Balaguer 2022).

A digitised constitutional law that, while normatively capturing the digital by incorporating the axiology of constitutional rights into the design and implementation of new technologies, apprehends the form and mode of digital production by generating a legal rationality that is superimposed on the algorithmic.

The aim would thus be to articulate a normative state response to the constitutional impasse generated by the technological impulse of the political economy of the digital world, recovering the function of the Constitution as a structural and structuring space for legal relations between public authorities and BigTech. In any case, the theorisation described above is aware of the spatial shortcomings of the fundamental national texts for the legal organisation of private actors that transcend the territorial spaces where such texts deploy their normative nature and scope. Hence, these attempts to constitutionalise the digital by national rights have been surpassed by other proposals of global scope which have their most complete formulation in the construct of digital constitutionalism.

The differences with the previous approaches are that, while the proposals of domestic law advocate limiting the constitutional emptying from the State-private power relations; the theses of digital constitutionalism are aimed at articulating a middle way between the regulation of the digital challenge oriented towards the market and the regulation of such a challenge from the orbit of state sovereignty (De Gregorio 2022, 290-296). In relation to the former, as opposed to a digital capitalism that predetermines its rule of tech, there is an intersection with the rule of law of a digital humanism where the duality of objectives converge through a set of principles and values centred on human dignity that does not imply an intervention in the digital market (De Gregorio 2023, 22), in accordance with the redistributive logic of social constitutionalism, but rather an anchoring of the legal form and mode in the actions of private operators in this market, an indirect regulation, closer to the ordoliberal model. This intersection would be present in the AI Act, where the guarantee of a digital single market in harmony with the digital political economy and free competition converges with the indivisible and universal value of human dignity, as enshrined in the Charter of Fundamental Rights of the European Union.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

Nonetheless, I believe that this approach to AI tangentially borders on two essential questions: why do digital social relations acquire a legal character, and why do they require legal mediation? A priori these questions may seem simplistic, even meaningless, because if it has been argued that AI produces effects on rights in their individual and collective dimension, the application of constitutional normative logic cauterising potential AI-rights conflicts follows. However, if we reduce Law to purely normative legality, we are excluding from the analysis the social conditions that make the efficacy of the legal form possible. In other words, we are avoiding why the constitution ascribed to a specific form of state that emerged after the Second World War continues to be a guarantee of normalisation as a legal process and project despite spatial limitations (technological operators act on a global scale) and, we add, of legal technique (is it possible to speak of digital rights from a theorisation of rights designed for disputes between public authorities and individuals -whether natural or legal-)?

The theorists of digital constitutionalism themselves develop their reflections around the spatial insufficiencies of the state and the need to articulate a normative legal response on a supranational scale. However, they then advocate reproducing the state constitutional legal arsenal for dealing with digital phenomenology in the supranational framework. The state political framework would be insufficient, but not the constitutional legal framework. These considerations ignore two decisive elements in the analysis of financialised capitalism in its digital phase: the unquestionable protagonism of states through their political and legal actions to guarantee the project of global market power in its financial form and its current digital process; and, consequently, the inseparable relationship between digital power - state and Law.

According to this methodological approach, Law, the Constitution, is not abstract Law. Fundamentally because I understand that the Constitution cannot be shaped as a stony text independent of the concrete content of the legal norms, in the sense that it retains its meaning, even if this concrete material content varies. Constitutional normativity, its formal and material supra-legality, is linked to the effects of the force of form, understood not as a legal formalism that positivises an artificially constructed legal order, but as a legal materialism linked to the social foundations of the force of form. In other words, the singularity of the Theory of the Constitution should not consist of limiting itself to the mere description and formal and logical analysis of the norms, but should explain according to what interests the norms have been produced, what meaning the relations

they regulate have in reality and what are the real forces that guarantee their application in practice (De Cabo 1993, 271).

A methodological approach of the kind described above is to seek a materialistic explanation of legal regulation, according to which Law expresses a specific socio-economic relationship before it is a norm, and as such must be investigated (Pašukanis 1976, 73-74). Hence, the inclusion of the social conditions that make the efficacy of the legal form possible makes it possible to understand the structure of the power relations present. In this construct, the insertion of constitutionalism in the capitalist mode of production and its structural relation constitutionalism-capitalism, the fundamental thing is to discover the relations with these structural elements.

2.1. *Digitalisation: the liquefaction of analogue legal orders? Or a new constitutional paradigm?*

According to this materialist approach to juridical regulation, social and economic relations, the social and the economic, are related through the state. For this reason, we speak of the Constitution in the form of the liberal State, of the Constitution in the form of the social State, and, although still a minority thesis, of the Constitution in the form of the market State (Maestro 2015, 53-94). In this form of market state, legally mediated by its Law, the power of the state is financialised, assuming functions of reordering politics and the economy, of recomposing the subjects of the conflict and the limits of state power. In this association, the Law is the embodiment of this correlation of social forces and the intensity of the conflict between them. The Law therefore acts as the structure linked to the conditions of social reproduction that make the effectiveness of the legal form possible. From this point of view, it is possible to see that, if the liberal state and the social state had their Law, which corresponded to the consecration of the political-economic division in the liberal state, and to the capital-labour pact in the social state, it can be inferred that the market state also demands a form of power and Law, which makes explicit the conditioning factors of the new conditions of the political-economic relationship.

It is therefore pertinent to take as a reference point the context that determined the reformulation of capitalism in the second post-war period, its reconfiguration under the global project of the financialisation of economies based on the unconditional centrality of the market, its correlate of free competition and the depoliticisation of

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

economies. It is at this point of inflection of the old order and the emergence of a new one that the Gordian knot of the type of power, state and Law necessary for the reordering of the new foundations of social reproduction is condensed, where, as it has already mentioned, the stagnation of the rate of profit was the main trigger.

In relation to the Constitution, its function of social integration is limited through the translation of the decisions of constitutional systematics to procedural and regulatory rules, to the hermeneutics of constitutional justice and legislation, projecting a function of normativity understood in an ahistorical sense, which makes it possible to reconcile the moment of constitutional rupture without compromising its formalist validity or formal identity (García Herrera 2022, 239). The political Constitution assumes the changes underway, distancing itself from the legal Constitution, which acquires a tautological sense of indefinite validity, while its validity and efficacy is transited in the political and jurisdictional arena that endorses the new market order. Only in this way is it possible to conceive of a constitutional law whose functions of limiting power and guaranteeing rights are predicated without delving into the effects that the new social conditions of reproduction have on these functions. The interpretation of the causes as an external link to the national constitutional order (the global power of the globalising market) makes it possible to retain the affirmation of constitutional normativity, as it cauterises the interpretation of the causes from its consideration as an internal link that logically affects the reflection of the function of Law and the correspondence with the new state functions (García Herrera 2021, 109).

From the new function of global market power, consisting in undoing the function of the normative system of social constitutionalism (not of the aseptic constitutional state) of political direction of capital accumulation under the juridical mediation of the social integration of conflict, derive effects that affect the structure of the Constitution. The new function determines the material relativism of the constitutional order, which is now legitimised on the basis of guaranteeing terms that allow a formal articulation of the order, avoiding the confrontation between the material bases of social constitutionalism and the new market constitutionalism. The production of the legal norm is redirected to the opening of the constitutional contents that are not refractory to the redefinition of its contents, given the rupture of the global project of social constitutionalism that linked and founded the written constitution of the social state form. This opening is determined by the fact that the

production of the legal norm is now defined by the new needs of the market economic system (García Herrera 2015, 143-144).

The new relationship that is created between the State and the market, parallel to the new relationship established between the economy and politics presided over by the centrality of the market, supposes that the modes of production come from the ends assumed by the new form of market state. Hence, we are not dealing with a specific constitutional mutation, but with a material constitutional rupture, because the rupture of the material bases that legitimised the form of the social state and its constitutional law alter the function of the Law of social constitutionalism, affecting its structure and the form of the social State itself. The change in the functions of the State, from direction to management of the economic processes, through a system of economic and financial links to political power, defines the new material constitution (the new material conditions) by materialising the transformations that the new form of market State incorporates (Maestro 2022, 182-185). This loss of validity and legitimacy of the material constitution of the social state form, far from being settled with a crisis of validity and legitimacy of the formal constitution, has been resolved by situating the privatisation of power in the structure of the global market form, as a space absent of controls, and internalising the material constitution of the global market form in supranational (EU) and state spaces as spaces of control and guarantee of the constitutional order of the market.

The constitution-industrial capitalism relationship was juridified in the constitutional device through the constitutionalisation of the redistributive conflict in the form of Fordism. For its part, financialised capitalism, consisting of a process of recovery of the return on capital after a period of decline in the rate of profit due to the crisis of the 1970s and 1980s, is formalised through legal legality, incorporating an insuperable contradiction between state democracy and the markets, where the former is incapable of deciding on the conditions of life, on the foundations of social reproduction. In order to overcome this contradiction, the coexistence of systems-orders is called for, which are concretised at different levels of action, the state and, in our case, the European legal order. The latter legalises the new functions assigned by global market power to supranational and state spaces.

The deregulation of capital, the renunciation of fiscal progressivity, the dissociation of the market from the social interest, are the function of the new structures of state and supranational power. States become functional to reinforce the logic of subordination to the market externally and internally. Internally, through fiscal discipline to guarantee

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

the macroeconomic balances necessary for the protection of the unconditional market against any redistributive intervention. Externally, states are articulated as globalising agents around projects that favour the global market. In this sense, although the EU can be interpreted as a supranational space that supports member states' capitals by improving the conditions of their competition in a global space, this approach cannot be attributed to a reconfiguration of the relations between politics and the market characterised by the social protection of member countries in the face of potential negative externalities of globalisation (Maestro 2011, 170-71). Basically, because the supranational space affirms the globalising strategy by organising the set of social relations around the centrality of the market. This implies the absence of any redistributive dimension and the connection of potential public spending policies to contexts of crisis of global market power, as experienced during the management of the financial crisis.

The crisis of financial capitalism in 2008 converges precisely with the advance of the dynamics of the digitalisation process aimed at safeguarding a strategy of financialised accumulation which, as a result of its own self-induced crisis, is incapable of expanding and therefore needs to find new spaces to reproduce itself through so-called immaterial capital (Rifkin 2001, 41). The uniqueness of this process of immateriality, however, is its concomitance with the material foundations of the market state: market centrality and autonomy. This implies that technical-scientific development does not exist as a denuded condition of the relations of production. In fact, capitalism has historically developed the productive forces through new scientific and technical innovations to overcome its crises.

The last century and a half have seen the second and third industrial revolutions. With the second, the modern factory system developed through Taylorism and Fordism, overtaken later, in the second half of the last century, by the progressive introduction of automation and digitalisation. The subsequent 'Toyotist model' laid the foundations of an industrial form based on 'just-in-time' production and aimed at increasing productivity by rationalising production. The progressive introduction of digitalisation in production, distribution and service processes has led to the coexistence of digitisation and interconnection processes over the years until the so-called 'fourth industrial revolution', whose main technologies include, among others, AI. Thus, technology is integrated into the economic and financial systems, reinforcing the guarantee of the telos of global market power, highlighting the conjunction of the evolution of technologies with the evolution of the configurations of economic and political power that we have pointed out.

2.2. *Analysis of the AI Act from the material-constitutional system of EU Law*

In line with the methodological approach developed, the AI Act is inserted in the European legal order that positivises the material bases of the market order. These material bases are institutionalised around weak governance (the disappearance from the European space of the institutionalisation of public intervention) and strong control (steering of economic processes as a guardian - guarantee of the autonomy of the market). Digital governance in accordance with these positivised material conditions forms the legal basis of the AI Act. In particular, Article 114, in conjunction with Article 26 of the Treaty on the Functioning of the Union, or the guarantee of the system decision on the functioning of the internal market. Precisely the function of European market constitutionalism, to avoid market fragmentation by consolidating its centrality, predetermines the chosen source of EU law, the Act, as its direct applicability and immediate vertical and horizontal legal effectiveness is syntonic with the legal mediation of the material bases of the order of social reproduction of digital capital. The permanence of the conditions of this order is produced through the technique of negative harmonisation of the (ordoliberal) order of the digital market (Farrand 2022, 112-116).

A functional approach where the fact that the objectives set out in national regulations are basically equivalent allows the elimination of obstacles to the free movement and provision of services even in the absence of harmonising provisions. This does not logically mean the withdrawal of Member States from the digital market, but the involvement of the state and its subordination to the same digital market rules that apply to private operators. A public-private inter-institutional relationship or cooperation between public and private actors 'in the AI ecosystem'. This last term, used by the AI Act in Article 58.2 f), seems to configure an organic digital sub-community within the organic-systemic community of the internal market, a presumed dialogic collaboration between public and private sectors.

However, it should be noted that the AI Act seeks to correct, by means of a uniform legal rule, the possible market failures of AI systems that arise from the existence of divergent national rules (intra and extra-Community). Thus, the configuration of this homogeneous legal framework takes place through the market guarantee rules on competition and the functioning of the common market, complementary in turn to the system of prohibitions derived from the economic freedoms that claim their definition from the unitary

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

instance, sanctioning the leading role of the EU in the imposition of the constitutional paradigm of the market. Specifically, "this Regulation ensures the free movement, cross-border, of AI-based goods and services, thus preventing Member States from imposing restrictions on the development, marketing and use of AI systems, unless explicitly authorised by this Regulation" (Recital 1 in fine of the AI Act).

Thus, the difficulties for the functioning of a perfect market for inter-regulatory competition in the EU Law have led to the introduction of a number of substantive principles or strengthening mechanisms such as the provisions concerning the free movement of goods and services, and the maintenance of competition or the principle of mutual recognition that are a response to the presence of obstacles to mobility ad intra AI systems. Thus, an attempt to imitate the market, in the sense of causing the results that would have been achieved if competition between regulations had been able to operate freely.

The political problem of regulation is resolved in the negative freedom from state control and political power, and is based on the cooperation of market participants through the market itself. In this market priority, the AI Act prohibits all those elements that are considered to restrict competition and thus hinder the development, market introduction, commissioning and use of AI systems. It is therefore the principle of cooperation, rather than subordination to the political link, that is, the distinguishing feature of the market constitutionalism model as it is implemented in the functioning of the digital single market.

Notwithstanding the above, it can be counter-argued that the precursor policy documents for the supranational regulation of AI link the legal-normative regulation of the digital transformation to a human-centred approach. Specifically, in the European Declaration on Digital Rights and Principles for the Digital Decade of 2023[9], the European Parliament, the Council and the European Commission set out the necessary adaptation of this transformation to the values and rights of the Union's order, which are established as determining factors from which to approach any legislative proposal on digitalisation in the European framework. It thus seems to be assumed that the new form of digital power can produce positive and negative effects in the legal sphere of the subjects that participate in the digital single market. An ambivalent nature that is projected in the AI Act when it warns (Recital 48) about the possible adverse consequences of

[9] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32023C0123(01)

an AI system (with special consideration for AI systems classified as high risk) for the fundamental rights protected by the Charter of Fundamental Rights of the Union (CFREU)[10], including

> the right to human dignity (Article 1), respect for private and family life (Article 7), the protection of personal data (Article 8), freedom of expression and information (Article 11), freedom of assembly and of association (Article 12), non-discrimination (Article 21), the right to education (Article 14), consumer protection (Article 38), workers' rights (Articles 27-28 and 31-31), the rights of persons with disabilities (Article 26), gender equality (Article 23), intellectual property rights (Article 17), the right to an effective remedy and to a fair trial (Article 47), the right of the defence and the presumption of innocence (Article 48), and the right to good administration (Article 41).

In the light of this supranational political consideration, we could point out that the digital single market deploys its horizontal effects on the participants in this market, which becomes a recognition of the effectiveness of the fundamental rights of the Charter *vis-à-vis* individuals (providers, suppliers, producers of services or goods on the basis of AI systems). To paraphrase Bilbao (2017), the fundamental rights of the Union also apply, ex Charter, to private-law relations. In fact, the fundamental rights that can be violated are located throughout the Charter's provisions, exhausting all the chapters that form its backbone: dignity, freedoms, equality, solidarity, citizenship and justice. The focus on the nature of the activity carried out, as opposed to the focus on the nature of the subjects, seems to transmute the once self-determination of private law to the asymmetries and inequalities that AI systems may entail, generating obligations to protect vulnerable subjects (Simoncini and Cremona 2022, 261, 263-264).

However, these theses of paradigm shifts between the categories of public and private law, blurring the boundaries of one and the other, which seem to mimic the virtuality of the single digital market, do not diminish how this subjectivity operates, acts within the framework of economic activities, free provision of services, free movement of goods, which circumscribe the potential vulnerability of the people to their status as consumers, even if their quality of vulnerability is emphasised. In this sense, the technique of the Union's

[10] https://www.europarl.europa.eu/charter/pdf/text_en.pdf

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

co-legislators has been to create a digital single market from the perspective of the consumer-subject, where vulnerability becomes the lack of transparency of service conditions, of terms of use, etc., but not the vulnerability generated by an unequal position not only in the digital single market, but in all spheres of society. The digital subject brought back to the rationality of the functioning of the market predetermined by rules or indirect regulation of the failures of the digital market for its correct functioning, is a subject decontextualised from the systemic, socio-economic, political and cultural inequality of the digital power. A subject whose digital juridicity becomes the antijuricity of structural material equality both in internal constitutionalism and in the constitutionalism of the Union, converging the detonating causes of such desubjectivisation as a social being.

2.3. *The antijuricity (?) of the «digital» subject*

The market state form also unfolds its legal effects in the relationship between market and rights, which leads us to analyse this question from the coordinates of the historicity of the legal subject, the institutional (state forms) and material (social order of reproduction) determinants of its emergence at each historical juncture. Basically, because the guarantee of rights is also a consequence of the legal form of social reproduction.

The impact of AI on fundamental rights has been present since the irruption of the digital market and the technological powers that guarantee the functions of cumulative logic in this market. The limitations to the exercise of techno-scientific capital, when it affects the constitutional dogmatic, is a constant in the analysis of the potential impact of AI systems on the so-called paradigm of the constitutional state. The reconduction of such parameters to constitutional dynamics, following a similar analogy to the categorisations of citizenship that are functional to models of social organisation, leads to the proposal of a mimesis of digital rights that would implicitly entail a digital citizenship. In any case, it has been warned of the difficulties of resorting to the traditional legal categories of limits to powers, not only because of the transnational scope of the companies that exercise technological dominance; and of guarantees of rights, since the object to be protected is dynamic given the continuously developing conformation of digitalisation (Balaguer 2022, 29-30). In short, the difficulties of an analogical Constitution called upon to regulate digital spaces (Castellanos 2023, 266).

However, in the descent into a taxonomy of such digital rights that either transform the contents of the traditional civil, political, economic and social rights of the so-called constitutional state in their connection with digital environments, or generate new rights, due to their singularity, such as neuro-rights (Reche 2024), which connect with the human neurocognitive sphere, paradoxically we continue to resort to the classifications of a general theory of rights whose universal validity is predicated, while pointing out its inadequacies (Castellanos, 2024, 271-300). In particular, the characterisation of the new digital rights as rights that connect with the material equality or objective dimension of rights do not warn about the material causes linked to a specific reality from which institutional and socio-economic conditioning factors derive (De Cabo 2001, 117-136).

In addition, and in relation to the objective dimension of the rights, the following clarification should be made: to the subjective dimension linked to the natural law idea of the individual, an objective dimension is added in the general theorisation of rights in the post-World War II period, which is substantiated, firstly, in an extension of the concept of freedom in the literal sense (negative freedom); to give way, secondly, to a content that transcends that established in the rights of freedom (positive freedom). In other words, it is about participation in political, economic and cultural life for the sake of the realisation of the principle of substantial equality, the axiological basis of which lies in human dignity.

However, the distinction between the objective and subjective dimension of the rights is not new, nor is it specific to the constitutionalism of the social state, because, together with other effects, the aforementioned objectification leads to the construction of the principles of the constitutional system on the basis of fundamental rights, which, considering the anti-statist and individualist imprint of the liberal state, implies spreading it throughout the constitutional and infra-constitutional system. This is connected with the denaturalisation of the transformations of the social state form through the systemic recourse to the constitutional state. It is therefore important to specify that, when I speak here of the objective dimension of rights in the framework of the social state form, which is not synonymous with the constitutional state, I am talking about realising the assumptions of the social state in rights, so that the aim is not to liberalise the system of rights, but to socialise it. To this end, rights must be extended from the sphere of the individual, of formal equality, to the sphere of inequality, which is the element of subjectivity of social constitutionalism.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

Returning to the assumptions of the historicity of the subject, and beginning with the institutional causes, these refer us to the individual-state relationship. If, as I have argued throughout this paper, we have witnessed a change in the way power relations between the state dimension and the economic dimension are articulated, establishing a new relationship between the state and the market, the impact on the subject and the rights of the social state form is total.

This nuance is not trivial, because characterising digital rights as benefit/welfare rights refers to one of the dimensions of social rights, that of the social reproduction order of Fordist capital, which acquired specific profiles in social constitutionalism. Specifically, the functional character of social rights to the regime of accumulation was articulated from a double perspective (Maestro 2017, 776-779).

Firstly, in terms of productivity and economic growth. The benefit character and economic content of social rights was presented as functional to economic dynamics, participating in the logic of the capitalist system. Not only did they not cost the economic system, but they contributed to its growth and expansion. In the redistributive process of the social state there was a coordination between the dynamics of demand and production. Through the socialisation of investment, the social reproduction of the labour force was promoted. In turn, through the link between the wage relation and the accumulation regime a virtuous circle was generated between the production capacities and the consumption progression of the working classes, which favoured the creation of wealth and its redistribution.

Secondly, social rights were presented as functional rights for the legitimisation of the system, in terms of social adherence and political stability. Thus, ensuring the enjoyment of social rights meant fulfilling the contents of the social-democratic pact represented by the social state, generalising welfare situations and reinforcing the legitimacy of the state. Social welfare rights contributed to maintaining the internal cohesion of the working class by preserving that capacity for mobilisation which allowed the conquest of full employment and which served to numerically increase the forces of the proletariat and to resolve the strategic struggle against capitalism on its own behalf. Double functionality, to the homogenisation of the working class and to the economic system, on which the flexibility of social constitutionalism and the weaknesses inherent in the redistributive pact were theorised.

From these coordinates, to predicate the benefit/welfare character of digital rights would mean situating them in the logics of the order of social reproduction of capital, and thus drawing a certain analogy

between the question of subjects and power in industrial capitalism and in financialised capitalism in its digital phase. In this regard, there has been an interesting debate from the perspective of the political economy of digital capital, that is, as a social relation of production that puts into operation certain characteristics of human nature, such as sociality and the ability to communicate, which are the foundations of companies such as Amazon, Google, Apple or Facebook, among other info-technological companies. The need to set limits to counteract this strategy of accumulation is shared, but differences arise when it comes to conceptualising how the digital economy works. For some, it is an accumulation by expropriation of intangible or immaterial goods such as knowledge (Zuboff 2020). For others, it is an accumulation by exploitation because the business model of infotech platforms is largely based on the production of a commodity, the result of the search –"real-time access to large amounts of human knowledge"- although they then offer it for "free" in order to sell advertisers selective access to their users (Mozorov 2022, 89-106).

In the first approach, privacy, digital anonymity, would be the element to be protected; in the second approach, the generation of new forms of free value should be remunerated through contractual relationships between users and technology companies, where the service provision would be bidirectional: the companies provide the infrastructure and the users provide the data structure from which the servers are fed to generate the information commodity.

But whatever form one or other form of accumulation, expropriation or exploitation takes, it remains embedded in the logics of capitalism. What changes radically is the relationship of the state to the strategy of capital accumulation, which is no longer oriented towards its direction, disciplining it, but towards its management, guaranteeing it. This is why the quality of provision must be inserted into the framework of this new relationship. A conclusion that we also draw from the EU Law.

The new political and legal decision places the market as the nuclear element that legitimises the new order to which it refers. In this sense, all those values connected with the centrality of the market are values that the political mode of being considers indispensable. The rights that reflect the values that accompany this new model are rights that are functional to the market's decision. The interests of these rights are not protected for their own sake, but as functional to mobilise the structures that serve the strengthening and functioning of the centrality of the market. This implies that the subjective dimension, which configured the original fundamental rights as rights of non-

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

interference of the public apparatus in the private sphere, is translated in the market order into a prohibition of the public authorities to distort the dynamics of the capital markets. And the objective right, typical of social constitutionalism, which extended the axial content of human dignity to the sphere of material deprivation, disappears. In its place it is installed other dimension that sees the establishment of the axial nucleus of the market, competition and competitiveness, as the only possible content of the right. In other words, it would be the right to participate in an open market and free competition for the sake of the realisation of the fundamental political decision whose axiological support lies precisely in the economic logic introduced by the culture of market constitutionalism.

However, this does not mean that the rights of the new order entail a recovery of the postulates that informed the liberal constitutions. Despite the structural analogy between the rights of the two models, insofar as they are articulated as rights of defence, it is not possible to establish continuity between the two formulas. Otherwise, the rights of the market order would come to represent a kind of mimesis of the rights of freedom, when in their construction and legal consequences, as we have just seen, their differences are notable.

In this constitutional model of the market, the market, not as a *locus naturalis*, but as a social institute, reflects the constitutional values inherent in the new constitutionalism. The prescriptive element does not lie in a harmonious composition of values and rights in principle of different signs, where free competition alternates with freedoms of expression, information, access to networks, highlighting the axial component of every aspect of social life; but in encouraging the conviction that the whole of society can function as a market.

This last aspect reflects the socio-economic causes or social order of reproduction. An order which is normativised and which, starting from the exclusive reference point of the market, is projected onto each and every space of the individual. On the one hand, the contents of the different claims in which the rights are articulated do not consist in the right to have for all demands a benefit from the public power (a social benefit right), but rather a competitive and open market, the single European digital market, because only through the market is it possible to increase general welfare by realising the most adequate satisfaction of the needs of the individual.

On the other hand, if the market is established as the preferred space for the realisation of citizens' demands, the latter acquire a subjectivity in accordance with the language and culture of the market.

The holder of rights is no longer the subject of the conflict on which the category of social rights was articulated, but the consumer, the user of the services provided or generated under the form of AI production. The transition from the subject of conflict to the subject of the market illustrates the formal dissociation of economic and social relations, although objectively linked to the inequalities of the subjects[11].

Thus, the vulnerabilities that in the social state represented the inequalities of redistributive conflict now take the form of formally juridified vulnerabilities, and are thus abstracted from the social force of digital power.

Final conclusions: the «constitutional» inconsistencies of digital constitutionalism

This reflection concludes by questioning, now at the European supranational level (although it should be pointed out that this autonomous approach is far removed from the pluralist approaches of multilevel cohabitation without systemic frictions between Internal Rights-EU), the thesis of the so-called digital constitutionalism of the Union which, roughly speaking, is based on the limits of the European digital single market outlined by the Charter of Rights of the Union, which, as has been pointed out, has human dignity as its backbone.

This thesis presupposes, as Terzis (2024, 14) observes, the power structures of technological corporations as a natural fact which, as such, must be modulated by the dynamics of digital constitutionalism, applying, as far as is of interest here, the narrative of the dogmatics of fundamental rights, without questioning that such structures have been generated by law, and not in the absence of law. That is, there is no constitutional normative vacuum in the framework of digital power that must be filled at the supranational level, in the absence of the competence of national laws, in order to limit the governance of private power in digital social interactions when these affect the structural principles of the national constitutional state (rule of law, democracy, rights).

In any case, there is a hollowing out of social constitutionalism from the logics of the form and mode of production of digital capitalism,

[11] "Data mining first creates statistical social groups, and then policymakers design tailored interventions for each segment of society. Tailor-made, individualised governance is more likely to exacerbate social divisions than to promote inclusion" (Eubanks 2021, 233).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

which is a different matter. And this decoupling that appeals to the rational logic of the normative and regulatory intervention of law is generated by what I have called the constitutional paradigm of the market. A paradigm that silences the asymmetries of power by deregulating the rules of political direction of the digital market and that generates its own axiology, principles and rights. On the basis of this observation, the following concluding reflections are derived, focusing on iusfundamentality as the alleged core of intangibility of digital supranational constitutionalism.

Firstly, it should be recalled once again that, just as it is not possible to consider the AI Act in isolation from the legal system on which it is based, neither is the EU Charter an autonomous text, but its analysis must also be carried out from a unitary perspective which places it in the legal, political and institutional context in which it has been developed. Its link with the material bases of the European integration project is what makes it possible to determine the true scope and meaning of the provisions it contains, of the values and objectives that inform it, and of the mechanisms envisaged for its effective action.

Secondly, the process of constructing the European system of fundamental rights has been carried out from the Union's own order and its sources (general principles), being the fundamental rights provided by the constitutions of the member countries a source of inspiration. Thus, the construction of rights from the aims and objectives of the EU Law connects them with the economic link, negative integration or centrality of the market. The market and the mechanisms that are articulated for its action are projected onto each and every one of the variables of the European order, and far from being configured as autonomous components, they are inserted into the gears of the market paradigm, constituting a virtuous circle that explores and exploits all the virtualities of the economic link.

Thirdly, fundamental rights participate in this inherent genetic of market constitutionalism expressing in their conceptualisation the values determined by it. Thus, the fundamental rights inherent in the Union's axiological code differ substantially from the classic idea of the fundamental rights of the individual. The subjects of these rights and the interests they protect are diverse, because the material bases on which the very idea of fundamentality is integrated are diverse.

Fourth, economic freedoms form the heritage of the most important 'fundamental' rights of the European order. The quotation marks are intended to highlight the functional use of fundamentality, which completely loses the meaning of the fundamental rights of the

constitutional state, where the status of a right as fundamental places it in a position of normative autonomy capable of conferring its own substantiality in its own right. On the contrary, in the European space, the fundamental status of a right is determined by its contribution to the market order, «freedom of consumption and freedom of economic activity must be 'felt' in the conscience of citizens as intangible fundamental rights (Demichelis 2018).

Finally, the market form becomes the social form, the socialisation from the market and its variables, in the one of interest here, the European digital single market as a technical and economic normative order that must be integrated into the Market Form, where the so-called digital constitutionalism can only be admitted from its convergence with the Market and its Law, as a complement to the competitive processes in a market society facilitating access and equality of opportunities (not of positions) to these processes. In this way, the market itself is the instance from which the vital (digital) needs of the Union or the Vitalpolitik of the 'Market Order' advocated by Rüstow (Kolev and Goldschmidt 2022, 453-460) are configured.

References

Balaguer, Francisco. 2022. *La Constitución del algoritmo*. Zaragoza: Fundación Manuel Giménez Abad.

Balaguer, Francisco, and María Luisa Balaguer. 2023. *Verdad e interpretación en la sociedad digital*. Pamplona: Aranzadi.

Berardi, Franco-Bifo. 2021. *La congiunzione*. Roma: Nero.

Betancourt, Michael. 2015. *The critique of digital capitalism: an analysis of the political economy of digital culture and technology*. New York: Punctum Books.

Bilbao, Juan María. 2017. «La consolidación dogmática y jurisprudencial de la drittwirkung: una visión de conjunto». *Anuario de la Facultad de Derecho de la Universidad Autónoma de Madrid* 1: 41-74.

Bradford, Anu. 2023. *Digital empires. The global battle to regulate technology*. New York: Oxford University Press.

Castellanos, Jorge. 2023. «Sobre los desafíos constitucionales ante el avance de la Inteligencia Artificial. Una perspectiva nacional y comparada». *Revista de Derecho Político* 118: 261-287.

Castellanos, Jorge. 2024. «Una reflexión acerca de la influencia de la inteligencia artificial en los derechos fundamentales». In *Ciencia de datos y perspectivas de inteligencia artificial*, edited by Francisca Ramón, 271-300. Valencia: Tirant Lo Blanch.

Chiariti, Massimo. 2021. *Incoscienza artificiale. Come fanno le macchine a prevedere per noi*. Roma: Luis University Press.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

Crawford, Kate. 2023. *Atlas de la IA*. Barcelona: Ned Ediciones.
Cristiniani, Nello. 2023. *La Scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo*. Bologna: Il Mulino.
De Cabo, Carlos. 2001. «El sujeto y sus derechos», *Teoría y Realidad Constitucional* 7: 117-136.
De Cabo, Carlos. 1993. *Teoría histórica del Estado y del Derecho Constitucional*. Vol. II. En *Estado y derecho en la transición al capitalismo y en su evolución: el desarrollo constitucional*. Barcelona: PPU.
De Gregorio, Giovanni. 2023. «The normative power of artificial intelligence», *Indiana Journal of Global Legal Studies* 55: 1-26.
De Gregorio, Giovanni. 2022. *Digital constitutionalism in Europe. Reframing rights and powers in the algorithmic society*. Cambridge: Cambridge University Press.
Demichelis, Lelio. 2018. «Ordoliberalismo 2.0 e ordopopulismo». *Economia e Politica. Rivista online di critica della economica politica*. Accessed March 23, 2024: https://www.economiaepolitica.it/_pdfs/pdf-8629.pdf
Esteve, José, and Javier Tejada. 2013. *Ciencia y derecho: la nueva división de poderes*. Madrid: Fundación Coloquio Jurídico Europeo.
Eubanks, Virginia. 2021. *La automatización de la desigualdad. Herramientas de la tecnología avanzada para supervisar y castigar a los pobres*. Madrid: Capitán Swing.
Farrand, Benjamin. 2022. «The ordoliberal internet? Continuity and change in the EU's approach to the governance of cyberspace». *European Law Open* 2: 106–127.
Finelli, Roberto. 2022. *Filosofia e tecnologia. Una via di uscita dalla mente digitale*. Torino: Rosenberg & Sellier.
Floridi, Luciano. 2017. *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*. Milano: Raffaello Cortina Editore.
García Herrera, Miguel Ángel. 2021. «Neoliberalismo y estado económico». In *Crisis de la constitución: globalización neoliberal e integración europea*, edited by Gonzalo Maestro et al., 39-135. Granada: Comares.
García Herrera, Miguel Ángel. 2015. «Estado económico y capitalismo financiarizado: propuestas para un constitucionalismo crítico». In *Constitucionalismo crítico. Liber amicorum Carlos de Cabo Martín*, edited by Miguel Ángel García Herrera et al., 137-242. Valencia: Tirant Lo Blanch.
García Herrera, Miguel Ángel. 2022. «Estado y soberanía en la nueva fase de acumulación: entre crisis de la integración europea y la reconstrucción del espacio global». In *La refundación de la Unión Europea y la nueva centralidad estatal*, edited by Ainhoa Lasa, Miguel Ángel García Herrera and Gonzalo Maestro, 225-276. Valencia: Tirant Lo Blanch.
Han, Byung-Chul. 2014. *En el enjambre*. Barcelona: Herder.
Kolev, Stefan, and Nils Goldschmidt. 2022.«Vitalpolitik». In *The Oxford Handbook of Ordoliberalism*, edited by Thomas Biebricher, Peter Nedergaard, Werner Bonefeld, 453–460, Oxford: Oxford University Press.
Lanier, Jaron. 2013. *Who owns the future?* New York: Simon & Schuster.

Maestro, Gonzalo. 2022. «Las precondiciones para la recuperación del espacio constitucional estatal». In *La refundación de la Unión Europea y la nueva centralidad estatal*, edited by Ainhoa Lasa, Miguel Ángel García Herrera, and Gonzalo Maestro, 176-185. Valencia: Tirant Lo Blanch.

Maestro, Gonzalo. 2017. «El Estado Social 40 años después: la desconstitucionalización del programa constitucional». *Revista de Derecho Político* 100: 769-798.

Maestro, Gonzalo. 2015. «Del estado social a la forma global de mercado». In *Constitucionalismo crítico. Liber amicorum Carlos de Cabo Martín*, edited by Miguel Ángel García Herrera et al., 53-94. Valencia: Tirant Lo Blanch.

Maestro, Gonzalo. 2011. «La globalización americana». *Teoria del diritto e dello Stato. Rivista Europea di Cultura e Scienza Giuridica* 1-2: 165-187.

Mitchell, Melanie. 2024. *Inteligencia artificial*. Madrid: Capitán Swing.

Morozov, Evgeny. 2022. «Critique of techno-feudal reason». *New Left Review* 133/134: 89-126.

Numerico, Teresa. 2021. *Big data e algoritmi. Prospettive critiche*. Roma: Carocci.

Pašukanis, Evgeny V. 1976. *Teoría General del derecho y marxismo*. Barcelona: Labor Universitaria.

Presno, Miguel Ángel. 2022. *Derechos fundamentales e inteligencia artificial*. Madrid: Marcial Pons.

Reche, Nuria. 2024. *Mens Iura Fundamentalia: La neurotecnología ante la Constitución*. A Coruña: Colex.

Rifkin, Jeremy. 2001. *The age of access: the new culture of hypercapitalism where all of life is a paid-for experience*. Los Angeles: TarcherPerigee.

Simoncini, Andrea, and Elia Cremona. 2022. «La AI fra pubblico e privato». *DPCE Online* 51 (1): 253-271.

Simoncini, Andrea, and Erik Longo. 2022. «Fundamental rights and the rule of law in the algorithmic society». In *Constitutional challenges in the algorithmic society*, edited by Hans-W. Micklitz et al., 27-41. Cambridge: Cambridge University Press.

Terzis, Petros. 2024. «Against digital constitutionalism», *European Law Open*, 1-17. doi:10.1017/elo.2024.15

Van Cleynenbreugel, Pieter. 2022. «EU by-design regulation in the algorithmic society: A promising way forward or constitutional nightmare in the making?». In *Constitutional challenges in the algorithmic society*, edited by Hans-W. Micklitz et al., 202-218. Cambridge: Cambridge University Press.

Zuboff, Shoshana. 2020. *La era del capitalismo de la vigilancia. La lucha por un futuro humano frente a las nuevas fronteras del poder*. Barcelona: Paidós.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 73-100
https://doi.org/10.18543/djhr.3189 • http://djhr.revistas.deusto.es/

Facing fundamental rights in the age of preventive *ex ante* AI: a contemporary form of discrimination

La encrucijada de los derechos fundamentales en la era del control *ex ante* asociado a la IA preventiva: Una nueva forma de discriminación

Mª Teresa García-Berrio Hernández
Universidad Complutense de Madrid. España
teresag-berrio@der.ucm.es
ORCiD: https://orcid.org/0000-0002-4205-4184

https://doi.org/10.18543/djhr.3191

Submission date: 06.06.2024
Approval date: 22.11.2024
E-published: December 2024

Citation / Cómo citar: García-Berrio, Mª Teresa. 2024. «Facing fundamental rights in the age of preventing *ex ante* AI: a contemporary form of discrimination.» *Deusto Journal of Human Rights*, n. 14: 101-125. https://doi.org/10.18543/djhr.3191

Summary: 1. Artificial Intelligence and human condition: for an ethical use of AI. 2. EU Artificial Intelligence Act: a new roadmap on fundamental rights risk management in the face of AI. 3. Preventive risk control: a contemporary form of discrimination. 4. Principle of Non-maleficence: Prevention of harm and preservation of human dignity in the face of the risk of AI. 5. Mitigating the discriminatory impact of biases in AI algorithms: seeking the beneficence principle. Conclusions. References.

Abstract: As Artificial Intelligence (AI) systems become increasingly integrated into the social fabric of contemporary communities, ethical considerations surrounding their impact on fundamental rights have come to the fore. Indeed, the growing significance of AI has recently prompted a pivotal discourse within academic and policy circles in Europe concerning the development of an ethical framework for human-centric AI. As part of a broader research project examining the implications of AI on fundamental rights, particularly the right to non-discrimination, our objective is to present a preliminary overview of fundamental rights' risk management in the context of AI. In light of the significant impact of AI on vulnerable individuals and minorities, our discussion will subsequently address critical areas of concern related to the EU AI Act, including algorithmic bias and its constituent elements of discrimination based on ethnicity or religion.

Keywords: AI, ethics, fundamental rights, algorithmic biases, EU policies

Resumen: A medida que los sistemas de Inteligencia Artificial se integran cada vez más en el tejido social de las comunidades contemporáneas, las consideraciones éticas en torno a su impacto sobre los derechos fundamentales cobran más fuerza. En este sentido, tanto en círculos académicos como políticos europeos se ha propagado en los últimos años un debate recurrente sobre la viabilidad de construir un marco ético favorable a una dimensión antropocéntrica de la IA. Como parte de un proyecto de investigación más amplio que examina las implicaciones de la IA sobre los derechos fundamentales, en particular el derecho a la no discriminación, nuestro objetivo es presentar una visión preliminar de la gestión del riesgo de los derechos fundamentales en el contexto de la IA. A la luz del significativo impacto de la IA sobre las personas vulnerables y las minorías, nuestro estudio abordará asimismo aquellas áreas críticas relacionadas con el Reglamento europeo de la IA, incluido el sesgo algorítmico y sus elementos constitutivos de discriminación por motivos étnicos o religiosos.

Palabras clave: Inteligencia artificial, ética, derechos fundamentales, sesgos algorítmicos, políticas europeas.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

1. **Artificial Intelligence and human condition: for an ethical use of AI**[1]

The implementation of Artificial Intelligence (AI) in our daily-basis routines represents an unprecedented anthropological disruption, with a direct impact on all natural, economic, and social structures of human communities. The advent of recent technological advances has posed a challenge to our ethical consciousness, particularly in light of the illusion of postmodern individual autonomy. This illusion is strongly marked by the high price paid in industrialized societies for the process known as "individualization", which has become a substantive marker of "reflexive modernity". Indeed, some of the most prominent contemporary sociologists and philosophers, such as Gilles Lipovetsky, Elizabeth Beck, and Zygmunt Bauman, have advocated for this perspective. Strongly influenced by the sociological tradition of Norbert Elias, in his work *Liquid Modernity*, Bauman conceptualizes the phenomenon of "individualization" as a process that transforms human identity "from a given into a task, burdening actors with responsibility for this task and for the consequences (and side effects) of their actions" (Bauman 2003, 20).

The term "individualization" thus refers to the social process that is the consequence of the strong development of individualism that characterized late modernity in the second half of the twentieth century –qualified in sociological terms such as the above– mentioned "reflexive modernity". Conversely, it is also the triumph of Libertarian logic during the first decades of the 21st century, which has led to the individual being regarded as the absolute owner and responsible for their own life. This process of reflection has rewarded the development of the capacity for self-determination above all else.

In the context of technological transformation of contemporary societies, sociologists like Ulrich Beck (2008) view "individualization" as a radical transformation of the personality structure of societies. In terms of Beck, this is because the isolated individuals are led to believe that we can be freed from the constraints of traditional societal structures, which enables us to have complete control over the development of our lives through the decisions we make in the processes of technological rationalization. However, the effect of this phenomenon of "individualization" is devastating: it results in the

[1] The present study is part of the MICINN "Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas" 2023-2025 (PID2022-136439OB-I00), supported by MCIN/AEI/10.13039/50110001103.

individual being detached from the community, weakening their trust in others until they are left unprotected in a virtual world in which it is increasingly difficult for people to develop in an autonomous way.

The extensive deployment of AI systems and digital media for surveillance and social control during the past few years has inevitably led us to face the "flip side of the coin", namely the risk associated with AI tools (Beck and Gernsheim 2003). Do we need to question, then, the veracity of the anthropocentric philosophical and ontological paradigm, which posits that the human being is the only being in the world endowed with consciousness and that, consequently, human beings are the only ones who deserve ethical treatment?

In view of the profound impact that developments in biotechnology, neuroscience, and disruptive technologies such as AI could have on our understanding of the world in the coming decades, an increasing number of scholars are advocating for an approach to ethics in the context of techno-sciences that prioritizes the hermeneutical perspective by emphasizing the interdependence of knowledge through the integration of emotional insights. In this alignment, we endorse the proposal of the Spanish philosopher Fernando Savater (2011) who calls for abandoning the scientific reductionism of positivism and Heideggerian phenomenological factualism. Instead, Savater advocates for the hermeneutic-critical approach, which is well-known among other philosophers such as Habermas, Apel, or Adela Cortina (2007) in Spain. This approach aims to reinterpret the reality of the human person vis-à-vis their vulnerability in relation to technology (Cortina 2011).

In light of the growing acknowledgement during late postmodernity of the vulnerability of nature under the yoke of human technological intervention, the use of the term "responsibility" marks the transition from "inescapable reciprocity" between fellow human beings –meaning "love thy neighbor as thyself"– to a "responsibility towards nature" of a markedly teleological character. This transition is opposed to ethics based on conviction. Indeed, the Kantian categorical imperative –to act in such a way that the principle of one's action becomes a universal law– could be adapted through new formulations of a collective imperative, through intergenerational action in the public sphere.

As observed by German philosopher Hans Jonas, it is accurate to conclude that AI will profoundly impact the world we live in. However, it will be ethical considerations that will ultimately shape the nature of this transformation. Indeed, the consequences derived from the use of disruptive technology exceed the traditional frame of ethics, forcing us

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

to question the "principle of responsibility" discussed by Jonas (1995) in his essay, *Ethics for Technological Civilization*. As this author notes, technological intervention has significantly altered the inescapable reciprocity aspect of the ethics of proximity by positioning nature at the service of humans, with profoundly detrimental and dehumanizing consequences. From Jonas' perspective, the newfound capacity for human action on the natural world has fundamentally transformed the very nature of ethics. From the moment that human beings are able to destroy, they are held to a new standard with regard to future generations: namely, "the responsibility for what is to come". This is the way in which the term "responsibility" is employed by Jonas, wherein responsibility is oriented towards the future and encompasses both the present and the past tenses.

The construction of a shared ethical framework for AI systems represents nevertheless a significant challenge, particularly in light of the postmodern phenomenon of ethical relativism. This challenges us to recognize that every community and individual may have different conceptions of what qualifies as morally acceptable. The collective approach previously outlined, in which Jonas' responsibility for the future is considered, could provide a more explicit justification for the assumption of a "universal moral paradigm". This is because it is based on shared human experiences and the enduring principles of philosophical traditions that have guided ethical grounding throughout history. Consequently, the question arises as to whether the construction of a shared ethical framework for AI is a utopian idea or rather an increasingly pressing need.

The establishment of a moral paradigm and a unified ethical framework for the advancement of AI is predicated on two fundamental assumptions. Firstly, as Fernando Savater notes, a universal ethical framework could serve as a safeguard against the shortcomings of the scientistic reduction of positivism and factualism. Secondly, it encourages a hermeneutic approach that promotes interdependence through the transversality of the emotional approach in human knowledge, particularly in the context of technology (Cortina 2011).

In addition to the above, the proposal of a universal ethical frame of reference in AI facilitates a collective understanding of the boundaries that must not be transgressed by AI, which aligns with the fundamental purpose of technology. The term "technology" has its etymological roots in the Greek word *τέχνη (téchnē)*, which signifies art, craft, or skill. From a broader perspective, technology is defined as a process or capacity to transform or combine existing elements in

order to create something new, thereby enabling the improvement and deepening of human existence. This approach to understanding technology is rooted in the Aristotelian concept that the value of goods, institutions, and social practices is contingent upon their intended purpose or end. The most accurate method for discerning the virtues –both ethical and dianoethical– appropriated to a process, craft, or skill, such as technology, when attempting to comprehend the *telos* of that process is the fundamental tenet of the Aristotelian Theory of Justice and the foundation of Virtue Ethics.

It is therefore imperative to engage in a comprehensive discussion regarding the ethical implications of technological advancements associated with the use of AI. This debate would entail a comprehensive examination of the potential harms and dangers that must be avoided, while also promoting those values of human interdependence that can establish an ecosystem of trust among citizens, stakeholders, and users in the face of a "potentially harmful use" of AI.

2. *EU Artificial Intelligence Act*: A new roadmap on fundamental rights risk management in the face of AI

For several years, the European Union has been engaging in a comprehensive strategy for responsible research and innovation in Techno-sciences, which is represented by the acronym RRI (Responsible Research and Innovation). The RRI program represents a novel approach to research governance, aiming to bridge the division between the scientific community and society. It encourages the socialization of techno-scientific environments, where civil society and technologists collaborate to align scientific research with societal values, needs, and expectations. More precisely, the RRI program encompasses six lines of action: (i) Citizen participation throughout the research process. (ii) Gender equality in work teams. (iii) Science education to improve educational processes and promote scientific vocations among the very young. (iv) Ethical awareness to foster scientific integrity, in order to prevent and avoid unacceptable research practices. (v) Free access to scientific information to improve open dialogue with society. And (vi) Governance agreements, with the aim of providing tools that foster shared responsibility among interest groups and institutions.

In this context, the EU strategy for an ethical and responsible program on research and innovation in techno-sciences has gained

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

particular prominence in recent years, particularly in view of the presentation of a harmonized system of rules within the EU in the field of Artificial intelligence, known as the *EU Artificial Intelligence Act* (EU AI Act). The EU AI Act represents a significant regulatory milestone in fostering a collaborative ecosystem of trust among citizens, stakeholders, and users in the context of AI, which has the potential to be employed in ways that may result harmful. It aims to establish a transnational AI regulatory framework, and it is the first cross-cutting legal regulation that is directly applicable in all EU Member States, eliminating the need for subsequent national transposition rules to be developed. Furthermore, the regulatory system established by the proposed EU AI Act is universal in scope, extending to all AI systems functioning as components of products or intended for placement within the European Union market, regardless of whether they are standalone AI systems or integrated components within larger products.

The initial aim of the EU AI Act was to control and manage risk by addressing deficiencies in existing legislation, with a view to establishing an effective risk-based approach to AI (Soriano 2021). Indeed, the proposal for a common European legal framework on AI incorporates a system based on risk management that establishes different information obligations for providers depending on the level of risk associated with the use of an AI system with respect to the guarantees of users' fundamental rights.

This distinction is reflected in the categorization of AI systems into three categories. The first level, AI systems of unacceptable risk (level A), is prohibited and applies to systems whose risk is so unacceptably high (see Title II). The second level, AI systems of high risk (level B), which is considered as high-risk systems, applies to systems that generate important risk or that could adversely affect the due guarantee and safeguarding of fundamental rights (see Title III). The third level, AI systems of limited risk (level C), applies to systems of limited risk that, though they are not considered high risk, have a series of transparency's requirements (see Title IV). There is also a fourth level for the remaining AI systems, which applies to all other permitted systems (see Title IX).

i. The proposed EU AI Regulation establishes a first minimum obligation for those AI systems considered to be low risk or limited risk (Level C). Specifically, these AI systems require a minimum level of transparency's requirements that allow users to make informed decisions under their consent. Therefore, we

would be dealing with a limited risk AI system when users are aware that the image, audio, or video content offered to them has been generated by an AI application or device. With regard to generative AI applications, such as ChatGPT, the EU AI Act proposes a special mention of the additional transparency requirements that must be met in order to be classified as "limited risk" applications. In particular, it imposes specific requirements for generative AI systems, including the following: (a) The content of AI system must be disclosed to the user as having been generated by an AI. (b) The AI system must publish periodic summaries of the copyrighted data used for training. (c) The AI system must prohibit the dissemination of illegal content.

ii. AI systems that could adversely affect security or the due guarantee and safeguarding of fundamental rights are considered in the EU AI Act as high-risk systems (Level B). The European AI Act distinguishes between two categories of "high-risk AI systems" for the purpose of this distinction. (ii.1) The first category includes AI systems that are used in products subject to EU consumer product safety legislation –such as toys, aviation, automobiles, medical devices, or elevators–. (ii.2) Secondly, high-risk AI systems are defined as those that enable the following activities: (a) biometric identification and categorization of natural persons, (b) management and operation of critical infrastructure, (c) education and vocational training, (d) employment, management of workers and access to self-employment, (e) access to and enjoyment of essential private services and public services and benefits, (f) management of migration, asylum and border control, and (h) assistance in legal interpretation and law enforcement.

iii. Finally, the EU AI Act considers those systems that pose a direct threat to individuals and to the guarantee of their human rights as unacceptable risk AI systems (Level A) and expressly prohibits them. This prohibition extends to three essential modalities of AI systems: (iii.1) AI systems that employ cognitive manipulation of the behavior of vulnerable individuals or groups, such as children and adolescents; this prohibition encompasses the potential for AI devices to encourage dangerous behaviors in children or to induce suicidal behaviors in adolescents. (iii.2) AI systems that utilize algorithms to generate identity biases for the purpose of classifying individuals based on their socioeconomic status or personal characteristics, including race, gender, nationality, sexual orientation, religion, etc. (iii.3) Finally, AI systems that use

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

biometric identification, both in real time and remotely, which employ facial recognition.

In this regard, the arduous parliamentary discussions that took place among the Expert groups during the legislative process of reflection on EU AI Act have resulted in the extension of the list of AI systems to be considered prohibited to five new modalities: (a) Real-time remote biometric identification systems, when performed in public access spaces that would allow mass surveillance. (b) Delayed remote biometric identification systems, with the sole exception that the use of such systems are performed by state security forces and corps for the prosecution of serious crimes and by prior judicial authorization. (c) Predictive AI system that are able to anticipate the risk of committing criminal or administrative offenses. (d) Predictive AI systems that enable the inference of the emotions of a natural person in the domains of law enforcement and border management, in workplaces, and in educational institutions. (e) AI systems that employ subliminal techniques to materially distort the behavior of the same.

Despite the substantial support that the EU AI Act is expected to receive in the upcoming years, one of the most contentious issues that has emerged during the legislative process of the EU AI Act is the proposal to impose an appropriate level of prohibition on those AI systems that pose an "unacceptable risk" to the fundamental right of non-discrimination and to the safeguarding of ethical principles against the consolidation of negative stereotypes about religious or ethnic minorities. Bearing in mind the potential for discriminatory outcomes associated with AI systems, it is indeed crucial to give appropriate consideration to the stipulations laid out in Article 5 of the EU AI Act.

In accordance with Article 5.1a) of the EU AI Act, the utilization of any AI system that employs subliminal techniques or manipulative or deceitful methods for the purpose of influencing the behavior of an individual or group of individuals –and which is not discernible to the individual– is explicitly identified as an unacceptable practice. In such circumstances, the capacity of the individual to make an informed decision is significantly constrained, thereby increasing the likelihood of significant harm being inflicted on the individual or on another person. Notwithstanding the above, the prohibition of an AI system employing subliminal techniques shall not apply to AI systems intended for therapeutic purposes, on the condition that informed consent is obtained from the patients themselves or, when appropriate, from their legal guardians.

The ethical implications of this regulatory clause –as introduced in the final version of Article 5.1 a) of the EU AI Act– are significant. Accordingly, any attempt to influence our deep or unconscious mental processes through the use of subliminal techniques, or any manipulative or deceptive techniques employed in AI devices for the purpose of influencing our decisions as users or consumers about what to purchase, what to consume, what to appreciate, or what to despise, should be banned and declared null and void.

This stipulation pertains to all AI systems that prompt individuals to make decisions otherwise unmade by those individuals themselves. This disposition thus simultaneously targets two distinct forms of influence: (i) the manipulation of decision-making processes and (ii) the dissemination of disinformation with the potential to alter ethical, moral, or ideological convictions or identity, as well as religious beliefs. The second effect, which pertains to the role of misinformation in influencing opinions that may alter convictions or beliefs, is a particularly salient issue that warrants comprehensive investigation. This is particularly the case given that the EU AI Act has adopted a framework whereby it falls on the aggrieved party to prove damages. Firstly, the legislation mandates the presentation of evidence indicating that the decision in question would not have been reached by the user in the absence of the AI system. Furthermore, the EU AI Act stipulates that proof of a risk of significant harm must be provided, although it does not provide a definition for this term.

It follows that, should one accept the proposition that the human unconscious is worthy of legal protection, it is not adequate to prohibit only those deliberately deceptive or manipulative subliminal techniques used by AI systems with the intention of making a profit. This is not only because they have a considerable impact on the ability to make an informed decision, but also –as described in Article 5.1.a) of the EU AI Act– because it prompts us to consider whether the concept of "own and voluntary act" is called into question. In the context of our study, it becomes evident that a legal framework is required to protect individuals from exploitation by those seeking to influence their actions below the level of conscious awareness. It thus follows that a framework of protection that is conducive to the human condition within AI systems is of great necessity.

An additional legal issue that presents analogous challenges pertains to the prohibition established in Article 5.1(b) of the EU AI Act. This article prohibits those AI systems that exploit the vulnerabilities of individuals based on their age, disability, or specific social or economic circumstances with the objective of materially distorting an individual's

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

behavior in a way that may cause them significant harm. The second provision of Article 5 has a considerable scope of application, encompassing AI systems that interact directly with users, such as chatbots or recommendation-based AI systems.

Moreover, identifying the areas of vulnerability that may bring an AI system within the scope of the prohibition set forth in Article 5.1(b) of the EU AI Act is also a challenging endeavor. This is because there is no definition of the term "vulnerability" in the Act itself or in relation to each of the characteristics listed in Provision 5.1 (b). Therefore, a broad interpretation would result in the prohibition of manipulative systems that exploit the vulnerabilities of specific groups with the intention of modifying their behavior and causing them harm or damage.

It is clear that this regulatory structure is inadequate because of the conceptual ambiguity surrounding the concept of vulnerability, which impedes the effective implementation of Provision 5.1(b). Furthermore, the legal approach places the burden of proof on vulnerable users, which is a further shortcoming. Firstly, it is essential to establish whether the primary objective of the AI system in question is to exploit the vulnerabilities of a specific demographic. Secondly, it is necessary to demonstrate that the AI system in question actually distorts the behavior of the aforementioned vulnerable individuals, rather than simply appearing to do so. In practical terms, an individual seeking to demonstrate that a particular AI system has deliberately exploited a specific vulnerability must address the considerable challenge of assembling a compelling body of evidence to prove the system's malicious intent.

In particular, the necessity of implementing specialized safeguarding measures for cognitive freedom in the context of generative AI systems becomes a pressing need. It is imperative that a commitment be made to develop a consensual international legislative framework that subordinates the design, production, and development of AI to the dignity of the human condition. Secondly, a comprehensive set of protections is required to safeguard the human condition. This necessitates the criminalization of any deliberately manipulative or deceptive techniques employed by the implementation of AI applications or devices with the objective of influencing the ethical and critical consciousness of individuals, thereby limiting their ability to make autonomous decisions and actions. Thirdly, it is imperative to acknowledge the necessity of recognizing the existence of a fundamental cognitive freedom, which is indispensable for the safeguarding of the unconscious mind. The unconscious mind is the

inalienable foundation of human individuality and, as such, requires protection. This would entail establishing the source of validity for the requirement of "free consent" in any legal act. The governance of AI systems is likely to become the exclusive domain of political and economic elites in the near future. This may well result in increased inequalities and the emergence of new mechanisms of social exclusion. It is therefore of the utmost importance that regulatory measures be put in place to prosecute and punish the use of AI as a device for social control and manipulation.

Furthermore, with the growing prevalence of automated and generative AI systems comes a concomitant increase in instances of algorithmic discrimination. In these cases, the use of AI perpetuates the vulnerabilities faced by specific disadvantaged groups and minorities. Indeed, the advancement and implementation of IA systems founded on the use of algorithms present a multitude of possibilities for the incorporation of biases that are detrimental to members of disadvantaged groups. One of the means through which algorithms may perpetuate the traditional discriminatory structures faced by minorities is through the selection and weighing of variables employed by IA systems for the measurement and prediction of the object under consideration.

The order of priority assigned to specific variables in the measurement of the phenomena predicted by algorithms can affect the outcome of these programs. To illustrate, if a bank's customer credit rating system prioritizes income level over savings capacity as an indicator of an individual's creditworthiness, this decision will result in a greater disadvantage for women, millennials, immigrants, and other vulnerable groups.

Humans may be particularly susceptible to the influence of AI due to the potential psychological harm caused by technology that seeks to exert control over human will. In recognition of the potential for exploitation of human vulnerability in contexts of technological disruption, the European AI Act acknowledges the necessity for accountability and protection of individuals in the context of AI-driven change. In light of these considerations, the EU AI Act prohibits the exploitation of vulnerabilities of groups of people on the basis of their age, disability, or social or economic situation. This is done in a manner that distorts their behavior and is likely to cause harm to them or to others. Furthermore, in the context of safeguarding fundamental rights to equality and non-discrimination, the EU AI Act strives to strike an appropriate balance between the advancement of individual autonomy and economic efficiency (Soriano 2023).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

3. **Preventive risk control: A contemporary form of discrimination**

It has to be noted that artificial intelligence systems are not immune to the biases and prejudices that exist in society. Such biases tend to pervade the algorithms themselves, thereby facilitating the propagation of discriminatory outcomes. Consequently, an unbalanced or inappropriate selection of data during the training of an AI system may result in the algorithm making unfair decisions that lead to the stigmatization of certain groups, minorities, and/or individuals. Such bias may be generated by a number of factors, including preconceived beliefs, predilections, or unconscious prejudices that have been acquired by individuals throughout their lives based on the sociocultural stereotypes they may have acquired at different stages of life.

The *Rome Convention for the Protection of Human Rights and Fundamental Freedoms* of November 4, 1950, contains in Article 14 an "anti-discrimination clause » which encompasses both a general equality clause and a clause prohibiting discrimination on certain specific grounds, including gender, race, and ethnic origin. Furthermore, the Council of Europe has endeavored to remove the limitations imposed by the current Article 14 of the Convention through the approval of Protocol 127, which recognizes a broad prohibition of discrimination. In particular, the first article of the Explanatory Report to Protocol 127 states: "The exercise of any right recognized by law shall be secured without any discrimination based, in particular, on sex, race, color, language, religion, political or other opinion, national or social origin, association with a national minority, wealth, birth or other status."

The European Court of Human Rights (HUDOC) defines the term discrimination in the case Willis v. United Kingdom (September 11th, 2002) as "treating differently, without objective and reasonable justification, persons in substantially similar situations". In a related case, Thlimmenos v. Greece (April 6th, 2000) the Court broadened the scope of this clause to encompass discrimination by indifference, which refers to the existence of discrimination when states do not treat differently, without objective and reasonable justification, persons whose situations are substantially different (McCrudden 2008, 712-724). In accordance with the HUDOC doctrine, there would be discrimination when individuals are treated identically under the law, yet their circumstances differ –that is, discrimination by differentiation– and when individuals are treated differently despite their comparable circumstances –that is, "discrimination by indifferentiation". Nevertheless, it seems unlikely that

this doctrine of "discrimination by indifferentiation" will have a significant future impact. However, examples of the latter can be found in the use of a type of algorithms used in AI applications for facial recognition, particularly *ante facto* predictive algorithms, which show a significant discriminatory bias when identifying people of different racial and ethnic origins (Berk et al. 2018, 1-24).

In light of the European legal system's foundation on liberal premises regarding the advancement of individual autonomy, the traditional European legal framework in the domain of equality and non-discrimination has historically integrated two distinct categories of legal instruments to safeguard individuals against discriminatory practices. (i) The initial type is a preventive or *pre-facto* instrument against discrimination. This is also known as an anti-classification legal instrument. (ii) The second type of legal instrument is a reactive or *post-facto* instrument. This is also known as an anti-subordination instrument (Ganti and Benito 2021).

The anti-classification legal instruments –also known such as *ante facto* prevention instruments– refer to those rules that prohibit the consideration of particularly suspect categories in decision-making processes. Consequently, these are reactive legal instruments in the face of situations of discrimination. In addition, anti-discrimination prohibitions function to some extent as preventive mechanisms, articulating mandates that seek to avoid discriminatory decision-making such as article 9 of the General EU Data Protection Act, which prohibits the processing of special categories of personal data, including racial origin, religious convictions, or political opinions. In contrast, the second type of anti-subordination legal instruments –also known such as *post-facto* repression instruments– seek to reverse those social structures that place persons belonging to certain groups or minorities in situations of disadvantage or discrimination. Consequently, these *post-facto* mechanisms aim to identify and redress all kinds of violations of the general non-discrimination principle.

The use of AI-based predictive algorithms in surveillance systems presents a significant challenge to the exercise of social control in the context of digital environments. Indeed, the entire community is placed under the dependence of a single criterion: *risk control.* The function of *ante facto* predictive algorithms is to determine the possible degree of risk posed by a person throughout the different stages of the criminal process, specifically in regard to the possibility of recidivism. From a purely normative perspective, the utilization of predictive risk algorithms presents significant challenges for legislators tasked with the legal regulation of long-term stability, the effective repression and

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

combating of violations of fundamental rights, and the pursuit of justice (Añón 2022, 17-49). Furthermore, it presents a challenge for legal professionals, as predictive AI systems are designed to anticipate human behavior, which could potentially result in the formation of discriminatory biases based on such factors as gender, nationality, ethnic origin, race, or religion.

A case closely aligned with the current topic of discussion is that of the Risk Indication System, which is also known by its acronym, SyRI System. This example will demonstrate the discriminatory implications of *ante facto* risk models. The SyRI System was used by the Dutch government to prevent and combat social security benefit fraud. This system allowed the Dutch public administration to use risk reports for claimants of child benefits in preventing the illegal obtaining of government funds in the field of social security. The SyRI System was established on the basis of the normative framework provided by national law, the so-called *Law on the Structure of Work and Income Enforcement Organization*, which contains in article 65.2 an extensive list of categories of information that may be processed in the SyRI system: in particular, gender, employment history, taxes, property ownership, education, health insurance, government permits, level of debt, track of public benefits received, and administrative sanctions such as traffic fines. To calculate potential evasion and fraud irregularities, the SyRI System algorithms linked all the applicants' personal data stored by government agencies and matched them with a "risk profile" generated from the information of other citizens with criminal records. Once any similarities and/or discrepancies were established, the system produced risk reports on a list of names as "potential fraudsters" that could be retained by the authorities for up to two years. Additionally, The SyRI System was substantiated in neighborhood projects in which government agencies identified those municipal districts most adequate to implement this risk assessment system: in practice, the poorest neighborhoods and municipal districts characterized by high rates of immigrant population. As a result, the Dutch administrative authorities wrongly accused hundreds of families receiving benefits of fraud simply because of their Moroccan or Arab origin.

This SyRI case prompted a landmark judicial precedent in Europe, which resulted in the first court decision to examine an algorithmic risk assessment system. The Netherlands Committee of Jurists for Human Rights v. State of the Netherlands is a judgment issued on March 6, 2020, in which the court concluded that the SyRI system had not only affected the human right to privacy, but also violated the transparency

requirement of Article 8 of the European Convention on Human Rights. Moreover, the court examined the legitimacy of the government's use of citizens' risk reports to determine the allocation of social benefits. It concluded that the SyRI system was "neither transparent nor verifiable" not only because such a system could be used to create data profiles of individuals for other purposes, which are prohibited by law, but also because the risk models used by the Dutch government were never published. The interested parties were not notified in advance of the above when their data were entered into the SyRI system for the preparation of their risk profile before the public administration. Indeed, with regard to the balancing test, the court determined that a risk report has a non-negligible legal effect on the right to privacy of the individual subjected to algorithmic scrutiny, because such a report cannot preclude the use of sensitive information in subsequent procedures and communications between citizens and the public administration. Based on this reasoning, the court dismissed the "declared interest of the Dutch government".

Nevertheless, it should be acknowledged that when algorithms implement discriminatory practices based on so-called "suspect categories" or "*ante facto* prevention categories", they often employ a non-maleficence approach, whereby ostensibly impartial measurement criteria are, in reality, utilized in a manner that ultimately results in the disadvantage of individuals belonging to ethnic and racial minorities when compared to their non-minority counterparts. We shall now proceed to provide further clarification on this matter.

4. **The principle of *Non-maleficence*: Prevention of harm and preservation of human dignity in the face of the risk of AI**

In 1979, two distinguished American philosophers, Tom Beauchamp and James Childress, published a seminal work entitled *Principles of Biomedical Ethics*, which laid the groundwork for contemporary discourse within the field of ethics applied to medical sciences. In this publication, the aforementioned philosophers put forth four ethical principles as follows: (i) respect for autonomy, (ii) the principle of non-maleficence, (iii) the principle of beneficence, and (iv) justice. The authors presented these four principles, which have long been observed in human societies and have governed ethical behavior, as applicable to any culture or society (Beauchamp and Childress 1994).

The principle of non-maleficence is rooted in the classical medical maxim *primum non nocere*, which can be translated to "first do no

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

harm". It refers to the ethical obligation of avoiding any intentional infliction of harm. Indeed, the principle of non-maleficence can be defined as the obligation not to cause harm or to prevent harm from occurring. It encompasses the prohibition against killing, inflicting pain or suffering, and causing disability. Such a breach constitutes a public wrongdoing and is therefore subject to legal consequences.

Moreover, there is a clear distinction between the principle of not inflicting harm upon others, which encompasses behaviors such as theft and murder and the obligation of beneficence, which aims to safeguard personal interests or advance the collective good. When applied to AI systems, the principle of non-maleficence would ensure that such systems prioritize the safety of individuals or prospective users, as well as the preservation of human dignity. Consequently, this principle would serve to reduce risk and enhance transparency and explainability. More precisely, there are numerous instances in which AI devices have already incorporated the principle of non-maleficence with the objective of enhancing user safety. A case that exemplifies this concept can be observed in the automotive industry, particularly in the integration of AI systems into autonomous vehicles with the objective of reducing traffic accidents and enhancing road safety. The deployment of autonomous vehicles has the potential to result in a significant reduction in accidents caused by human error, which include driver inattention, visual fatigue, and lack of reflexes. Given the goal of AI devices applied to autonomous vehicles of reducing potential harm while simultaneously maximizing the safety of individuals on the road, such an approach would thus represent an instance of the implementation of the principle of non-maleficence.

In contemplating the possible applications of the non-maleficence principle to AI systems, it is *prima facie* necessary to examine the role of virtues such as kindness, empathy and compassion in the machine learning of AI systems, in natural language processing for the design of AI applications that are able to perceive, understand and respond to human emotions. Indeed, in the process of machine learning, AI systems are capable of processing vast quantities of data in order to make predictions about human behaviors that have already occurred. It is not thus simply a matter of creating intelligent machines that can replace humans in their reasoning and cognitive capacities. Instead, it is about fostering virtuous AI that reflects the best of human moral aspirations.

Consequently, if AI systems are trained to collect data that exemplifies the essential virtues to the human condition –such as kindness, empathy, solidarity, courage, prudence, and compassion– they can thus be enhanced with the ability to recognize and respond to

situations in ways that promote the common good in human communities. Machine learning offers an endless array of possibilities for instructing AI in the moral obligation to prevent or alleviate harm –therefore, to do good– and in the duty to help others over and above private interests. In other words, to act for the greatest possible benefit, seeking the greatest possible general welfare.

5. **Mitigating the discriminatory impact of biases in AI algorithms: Seeking the beneficence principle**

The term "beneficence" is generally accepted as the act of performing benevolent acts or actions that are perceived to be beneficial to others. Beyond the necessity to abstain from causing harm to others, the principle of beneficence obligates individuals to demonstrate concern for, and actively promote the well-being of, those around them. Indeed, the term "beneficence" is generally understood to encompass a broad range of behaviors, including acts of mercy, kindness, charity, altruism, love, and humanity.

Moreover, as defined by Beauchamp (2003, 12), beneficence encourages individuals and institutions to feel an ethical obligation to contribute actively to the welfare of the community by promoting civic virtues such as altruism, solidarity, compassion, and social responsibility in human actions. This principle implies a beneficial action that prevents or counteracts evil or harm, and additionally confirms the absence of acts that could cause harm.

The principle of beneficence furthermore represents a fundamental tenet within ethical theories, including the moral doctrine of utilitarianism. This is evident in the formulation of the utility principle, which states that actions should be taken to promote the welfare and act in a way that maximizes the happiness of the greatest possible number of people. This approach to the shortcomings of utilitarian reasoning is particularly evident in the context of AI, as it necessitates the focalization of efforts on mechanisms of beneficence that allow for the mitigation of unnecessary harms associated with AI systems, especially those that could significantly compromise collective welfare. It is therefore imperative that any AI system developers address the issue of mitigating the impact of biases in AI algorithms in order to comply with the principle of beneficence. Let us elaborate further on the concept of beneficence, which enables the mitigation of unwarranted damage caused by AI systems, particularly those with the potential to significantly harm collective welfare.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

There are no *ex ante* regulatory control mechanisms to ensure that AI systems are not discriminatory. Indeed, one of the primary limitations of the European legal framework is that mechanisms against discrimination typically operate *ex post*, that is, after the discriminatory action has already occurred. From a purely normative perspective, the implementation of *ante facto* predictive AI systems presents significant challenges for the legislature in its efforts to effectively repress situations that violate fundamental rights (Gerard and Xenidis 2021).

Algorithmic discrimination may also originate from errors or biases present in the databases utilized in the development of automated decision-making systems, as AI systems using predictive algorithms are designed with data related to the phenomenon they seek to predict. Once the system has been trained, its performance will be evaluated with data used to detect its level of accuracy. For instance, a database of arrests and convictions may contain primarily data on individuals from ethnic and racial minorities as a consequence of the pervasive discrimination they have historically faced in their interactions with law enforcement and the justice system. In this instance, the algorithm would be encouraged to learn that certain persons belonging to certain ethnic or racial minorities are more likely to engage in criminal activities. In other cases, the use of algorithms may serve to perpetuate stereotypes that underpin social structures of discrimination (Makonnen 2007). For example, the results yielded by entering specific combinations of words into Internet search engines, such as Google, have been found to reproduce gender roles or at least to contribute to the consolidation of negative stereotypes about religious or ethnic minorities (García-Berrio 2023). If the data used to train an AI system is of poor quality, the result may be that the algorithms induce us to undertake decisions that result in stigmatization of certain groups or minorities. This is due to the so-called "biases", which are understood to be preconceived beliefs, predilections, or unconscious prejudices that have been acquired throughout one's lifetime based on the sociocultural stereotypes with which one has been educated (Castellanos 2023).

Furthermore, predictive AI systems pose a challenge for legal professionals, as they are designed to identify patterns in human behavior. As we have pointed out, this may lead to the creation of discriminatory biases based on factors such as gender, nationality, ethnic origin, race, or religion. In light of the historical structures of discrimination that have placed certain ethnic groups and religious minorities in positions of disadvantage or subordination, it is evident

that when an AI system employs *ante facto* predictive algorithms, the validation and test data utilized to train the AI system would probably reflect historical structures of discrimination based on race, gender, religion, etc. As a result, the system may assume that the biases it contains are accurate or valid. It has, in fact, been demonstrated that the current bias produced by the use of AI systems is due to the imbalanced representation of ethnical traits that developers of AI systems employ in the training data (Žliobaité and Custers 2016). This representation will tend to include, for instance, a greater number of male and light-skinned faces. Conversely, if AI systems are employed as a predictive tool to generate profiles of potential perpetrators of a homicide, the validation data set will contain information related to homicides that have already been solved. Consequently, male and dark-skinned faces would predominate. For instance, a series of predictive risk algorithms have been implemented in recent years that can be applied to persons who respond to criminal stereotypes associated with different racial groups or ethnic origins. These stereotypes may increase the perception of guilt. This was exemplified by the long-standing use of the COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions)* model in the United States to assess the likelihood of recidivism among criminal justice system participants. The COMPAS system revealed overt algorithmic discrimination against African-American males, who represent the majority of U.S. prison population with the longest prison sentences –including life imprisonment– of any other racial and gender combination. The COMPAS system was found to exhibit a pronounced racial bias as far as African-American males being more likely to be misclassified as higher risk –a profile that was reflected in their COMPAS scores–. This flagrant racial bias in the COMPAS system has received considerable public attention, prompting concerns about the potential discriminatory impact of AI algorithms when used in jurisdictional decisions in the criminal justice system.

In light of the discriminatory impact of biases in predictive AI algorithms, article 44 of EU AI Act outlines a number of mandatory requirements to be met by the training, validation, and test data of high-risk AI systems, as well as by the individuals or organizations responsible for collecting such data and processing it. Article 44 of the proposed EU AI Act states that "high data quality is essential for the performance of many AI systems, especially when techniques involving the training of models are used". The objective is to ensure that the high-risk AI system performs as intended and safely, and that it does not become the source of discrimination prohibited by EU AI Act

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

(Hacker 2018). In addition to prior regulation, the second paragraph of Article 10 states that "training, validation, and testing data sets shall be subject to appropriate data governance and management practices". The fourth paragraph of Article 10 continues to elaborate the above: "Training, validation, and testing data sets must be considered in accordance with the intended purpose, taking into account the specific geographical, behavioral, and functional characteristics of the environment in which the high-risk AI system is intended to be used".

Despite the multitude of challenges to the principle of beneficence in the context of potential discrimination due to algorithmic bias, there is still a glimmer of hope. In light of the above, it is encouraging to report that a new guarantee of "human supervision" has been introduced in the EU AI Act. This new measure requires those responsible for the management of AI systems to be aware of the risks associated with bias, automation, or confirmation of the potential issues inherent in this type of digital application. In this regard, the European Parliament calls upon managers of AI systems to comply with their legal obligation to provide specifications for the input data or any other relevant information regarding the data sets used in AI systems, taking into account the intended purpose and the reasonably foreseeable misuse of the system.

Indeed, the EU AI Act introduces a new ethical duty favoring a recognition of the pivotal role played by intersubjectivity and the human condition within AI systems. This allows us to highlight the main disadvantage of AI: artificial intelligence and its algorithms lack the capacity to feel and possess no moral conscience. They are capable of understanding, but not of comprehending.

Conclusions

Increasing affective learning in automated AI processes serves to augment the capacity of AI systems to discern, comprehend, and respond to the nuances of human emotion. Nevertheless, as has been discussed in this paper, it is imperative to exercise caution in the acceptance of such advances. While AI systems may demonstrate empathic, kind, and compassionate behavior, they certainly lack the emotional connection that derives from human experience. This can be articulated in Kantian terms as the condition of humanity.

Any effort to establish an ethical framework for AI has the potential to imbue technology with a humanizing quality through the

promotion of virtues intrinsic to the human condition. This notable dedication to incorporating human factors into AI processes highlights the pressing concerns associated with the integration of AI in our daily lives. Consequently, integrating the three ethical virtues of kindness, empathy, and compassion into the configuration of AI systems paves the way for the creation of AI with a profound sense of humanity.

As our research illustrates, the ethical quality of AI systems can be enhanced through the implementation of fairness, which in turn facilitates the acknowledgement of the principle of non-maleficence. This is merely a method of circumventing the potential detriment that may result from the implementation of algorithmic biases. In addition, the promotion of empathy requires AI system developers to make use of the beneficence principle to offset hidden biases in *de facto* risk-avoidance algorithms, as well as discriminatory effects based on ethnic and racial identities that inevitably result in the segregation of minorities. In conclusion, the implementation of compassion into AI represents the pivotal impetus behind the mounting ethical pressure vis-à-vis the accountability of AI programmers and developers in the context of biased algorithmic sequences and the malevolent consequences of some of the latest generative AI utilities.

Consequently, any attempt to build an ethical framework for AI should acknowledge and accept the moral responsibility of human beings. The integration of kindness, empathy, and compassion into the design of AI systems would allow for the decisive prioritization of users' welfare, the promotion of fairness, transparency, and accountability, and the assurance that AI technologies serve the interests of citizens rather than those of technology corporations or *de facto* powers. Such a perspective should inform the development of AI algorithms that prioritize empathy, respect, and human dignity over the construction of discriminatory biases. Indeed, as we imbue machines with intelligence and decision-making capacity, the virtues we can instill in them become the very cornerstone of the ethical development of technology, especially in regard to addressing the potential systemic injustices that could result from a variety of biases in data and discriminatory algorithms.

In this study, we have selected a number of examples –including the SyRI system– with the intention of illustrating the primary challenge that the EU AI Act presents in terms of the automation of algorithms employed by predictive *ante facto* risk control systems. Indeed, the use of predictive AI systems by governments to generate "risk reports" for their citizens calls into question one of the epistemological foundations of the legal definition of the rule of law, namely the autonomy and self-determination of individuals.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

Those who adhere to contemporary interpretations of self-determination employ the Kantian ideal of moral autonomy to challenge the perspective of those who vehemently criticize the libertarian conception of personal autonomy as individualistic (or even selfish). In essence, Libertarians prioritize personal autonomy over subjectivities and preferences, a stance that is detrimental to the common good. For this reason, postmodern liberal thinkers such as Robert Young (1980, 573–576) and Joseph Raz (1986, 373) have proposed the notion of socialized autonomy, which effectively synthesizes the classical Kantian ideal of autonomy of the will with challenges such as those posed by the new applications of algorithms in AI systems. If our autonomy and ability to act freely are compromised through the use of predictive algorithms like the SyRI System, we no longer act according to a maxim that we have chosen for ourselves, but in compliance with a maxim that the community must establish for the common good.

Notwithstanding these limitations, it is crucial to acknowledge that one of the primary allures of AI applications is their capacity to present themselves as a means of overcoming human subjectivity, or even of eradicating stereotypes and social prejudices. The appeal to the certainty and neutrality of algorithms is an effective method for gaining acceptance and trust. Nevertheless, as illustrated by the SyRI System, there is a potential risk for algorithmic systems to be exploited by public agencies with the intention of establishing a repressive system that would be detrimental to public freedoms and fundamental rights, including the freedom of belief and the freedom of thought. Furthermore, the constitutional rights of citizens may be violated by the use of certain AI risk assessment systems, and human dignity may be infringed upon, particularly in the case of ethnic minorities and immigrant populations (Zuboff 2020). As previously observed, any individual subject to algorithmic scrutiny could potentially be prosecuted on the grounds of behavioral predictions generated by algorithms that may be perceived as risky. Rather than being prosecuted for the acts committed, individuals would be prosecuted *ex ante* based on identity biases associated with algorithms that take into account a range of factors, including an individual's level of income and indebtedness, their interactions on social networks, their place of residence, their religion, or their ethnic origin.

At the present time, the commendable attributes of AI systems are upheld on the basis of their reliability and predictability. However, it is crucial to consider the potential for predictive AI systems to be exploited for malevolent purposes, which could result in the consolidation of

authoritarian forms of governance and the undermining of democratic and citizen engagement. In this context, the construction of algorithmic patterns enabling machines to anticipate human behavior based on an *ex ante* predictability undermines the very concept of human autonomy. Accordingly, when adopting an ontological stance that recognizes the pivotal role of moral intersubjectivity and human autonomy, it becomes necessary to evaluate the primary limitation of predictive AI. This is because algorithms are unable to perceive human emotions and possess no moral conscience.

Current developments in the regulation of AI, as exemplified by the EU AI Act, posit human beings as the sole entity endowed with consciousness and the capacity to act autonomously. If we accept the proposition that autonomy and self-determination, in addition to the human conscience, are to be protected as a legal asset, we should consequently align ourselves with the European legislators. In that case, any recourse to predictive risk techniques employed by AI systems that have a significant impact on the aforementioned fundamental rights –including cognitive freedom, freedom of thought, belief, and religion– must be declared null and void.

References

Añón, María José. 2022. «Desigualdades algorítmicas: Conductas de alto riesgo para los derechos humanos.» *Derechos y Libertades* 47 (1):17-49.

Bauman, Zygmunt. 2003. *Modernidad Líquida.* Buenos aires: Fondo de Cultura Económica.

Beauchamp, Tom. 2003. «The nature of applied bioethics.» In *A Companion to applied ethics*, edited by Roger Frey & Christopher H. Wellman, 1-16. Malden: Blackwell Publishing.

Beauchamp, Tom and James Childress. 1994. *Principles of biomedical ethics.* New York: Oxford University Press.

Beck, Ulrich. 2008. *La sociedad del riesgo mundial: En busca de la seguridad perdida.* Barcelona: Paidós.

Beck, Ulrich and Elisabeth Gernsheim. 2003. *La individualización: El individualismo institucionalizado y sus consecuencias sociales y políticas.* Barcelona: Paidós.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns and Aaron Roth. 2018. «Fairness in criminal justice risk assessments: The state of the art.» *Sociological Methods and Research* 50 (1): 1-24.

Castellanos, Jorge. dir. 2023. *Inteligencia artificial y democracia: Garantías, límites constitucionales y perspectiva ética ante la transformación digital*. Barcelona: Atelier.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 101-125
https://doi.org/10.18543/djhr.3191 • http://djhr.revistas.deusto.es/

Cortina, Adela. 2007. *Ética de la razón cordial: Educar en la ciudadanía en el siglo XXI*. Madrid: Nobel.

Cortina, Adela. 2011. *Neuroética y neuropolítica: Sugerencias para la educación moral*. Madrid: Tecnos.

Ganty, Sarah and Juan Carlos Benito. 2021. *Expanding the list of protected grounds within anti-discrimination law in the EU.* Brussels: Equinet.

García-Berrio, Mª Teresa. 2023. «La sociedad digital como cultura del riesgo: Desafíos éticos e implicaciones legales del uso de sistemas de Inteligencia artificial para la evaluación de riesgos y la vigilancia preventiva.». In *Inteligencia artificial y Democracia: Garantías, límites constitucionales y perspectiva ética ante la transformación digital,* edited by Jorge Castellanos, 39-65. Barcelona: Atelier.

Gerards, Janneke and Raphaele Xenidis. 2021. *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law. European network of legal experts in gender equality and non-discrimination.* Luxembourg: European Union.

Hackers, Philipp. 2018. «Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law.» *Common Market Law Review* 55 (4): 1143-1186.

Jonas, Hans. 1995. *El principio de responsabilidad: Ensayo de una ética para la civilización tecnológica*, Barcelona: Herder.

McCrudden, Christopher. 2008. "Human dignity and judicial interpretation of human rights". *European Journal of International Law* 19 (4): 655–724. doi.org/10.1093/ejil/chn043

Makonnen, Timo. 2007. *Measuring Discrimination: Data collection and EU Equality Law*: *Thematic Report of the Group of Independent Experts*. Brussels: European Commission. Access December 9, 2024: https://www.tandis.odihr.pl/bitstream/20.500.12389/19825/1/03245.pdf

Raz, Joseph. 1986. *The morality of freedom*. Oxford: Clarendon Press.

Savater, Fernando. 2011. *Ética a Amador: Una invitación a vivir sin odio ni miedo.* Barcelona: Ariel.

Soriano, Alba. 2021. «La propuesta de Reglamento de Inteligencia Artificial de la Unión Europea y los sistemas de alto riesgo.» *Revista General de Derecho de los Sectores Regulados* 8 (1): 50-63.

Soriano, Alba. 2023. «Creando sistemas de Inteligencia Artificial no discriminatorios: Buscando el equilibrio entre la granularidad del código y la generalidad de las normas jurídicas". *IDP Revista De Internet, Derecho y Política* 38: 1-12. doi:10.7238/idp.v0i38.403794.

Young, Robert. 1980. «Autonomy and Socialization». *Mind* 89 (356): 565-576.

Žliobaité, Indre and Bart Custers. 2016. «Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models.» *Artificial Intelligence & Law* 24 (2): 183-201. doi: https://doi.org/10.1007/s10506-016-9182-5.

Zoboff, Shosana. 2020. *La era del capitalismo de la vigilancia: La lucha de un futuro humano frente a las nuevas fronteras del poder.* Barcelona: Paidós.

Opportunities and challenges of AI chatbots for digital youth information, advice, and counselling services in Europe

Oportunidades y retos de los chatbots de IA para los servicios digitales de información, asesoramiento y orientación para jóvenes en Europa

Alonso Escamilla
Universidad Católica de Ávila. Spain
alonso.escamilla@ucavila.es
ORCiD: https://orcid.org/0000-0003-0159-3576

Paula Gonzalo
Universidad de Salamanca. Spain
gonzalomoreno.paula@usal.es
ORCiD: https://orcid.org/0009-0001-0686-0162

https://doi.org/10.18543/djhr.3192

Submission date: 30.05.2024
Approval date: 22.10.2024
E-published: December 2024

Citation / Cómo citar: Escamilla, Alonso and Paula Gonzalo. 2024. «Opportunities and challenges of AI chatbots for digital youth information, advice, and counselling services in Europe.» *Deusto Journal of Human Rights*, n. 14: 127-154. https://doi.org/10.18543/djhr.3192

Summary: 1. Context of the problem. 2. Research objectives and methodology. 3. Evolution of youth information, advice, and counselling services in Europe. 4. The incorporation of artificial intelligence and chatbots in digital youth services. 4.1. Artificial intelligence. 4.2. Chatbots. 5. The perspectives of youth organizations. 5.1. What is a chatbot? 5.2. From giving information about a youth centre to getting involved in mental health issues. 5.3. Tailored for services and young people. 6. Opportunities and challenges of AI chatbots for digital youth services. 6.1. Generator of ideas and organizer of time. 6.2. Trust is earned through people and truthful information. 6.3. No time, no resources, no competences. 6.4. Who is after our data? Conclusions and recommendations. References.

Abstract: New technologies such as artificial intelligence (AI), applications and platforms are becoming more common in youth services and non-formal

education, with chatbots being key examples. However, many chatbots often fail to take into account the profiles, requirements and rights of young users leading to potential risks such as biases, polarization, and low data protection standards. In carrying out this research, a literature review was done to determine the history of youth services in Europe and the prevalence of chatbots. A series of interviews with representatives of organizations that either represented young people or provided youth services at the European level were held to share their experiences and describe the key features needed for a correct use of chatbots on youth services. This study highlights the practical possibilities and limitations of AI chatbots, and the need to co-design AI tools with youth organizations and young people in order to minimize threats and maximize the effectiveness of digital youth information, advice, and counselling services in Europe.

Keywords: Artificial intelligence, chatbots, youth work, Europe.

Resumen: Nuevas tecnologías, como las aplicaciones y plataformas de inteligencia artificial (IA), están volviéndose más comunes en los servicios juveniles y en la educación no formal, siendo los chatbots ejemplos clave. Sin embargo, muchos chatbots a menudo no tienen en cuenta los perfiles, requisitos y derechos de los jóvenes usuarios, lo que conlleva riesgos potenciales como sesgos, polarización y bajos estándares de protección de datos. En la realización de esta investigación, se llevó a cabo una revisión de la literatura para determinar la historia de los servicios juveniles en Europa y la prevalencia de los chatbots. Se realizaron una serie de entrevistas con representantes de organizaciones que o bien representaban a las personas jóvenes o proveen servicios juveniles a nivel europeo, para compartir sus experiencias y describir las características clave necesarias para un uso correcto de los chatbots en los servicios juveniles. Este estudio revela las posibilidades prácticas y las limitaciones de los chatbots de IA y la necesidad de diseñar conjuntamente herramientas de IA con las organizaciones juveniles y los jóvenes para minimizar las amenazas y maximizar la eficiencia de los servicios digitales de información, asesoramiento y orientación para jóvenes en Europa.

Palabras clave: Inteligencia artificial, chatbots, trabajo juvenil, Europa.

1. **Context of the problem**[1]

Digitalisation has transversely impacted and shaped the society we, as Europeans, know today. Simply, this transformative process has changed the way we carry out everyday tasks (Șerban et al. 2020). As Escamilla and Lonean (2021) describe, the COVID-19 pandemic unprecedentedly marked a before and after in the way we perceive digitalisation and the benefits and risks that new technologies, such as chatbots, pose to societies, impacting everything from education and social interactions to healthcare, governance, and, as this study will demonstrate, with particular implications for youth work. From this perspective, studying chatbots within youth work is essential, as they can serve as digital tools to enhance engagement, foster communication, and support the specific needs of young people navigating the evolving digital landscape.

These new technologies such as Artificial Intelligence (AI) have had a strong social and economic impact in the last decades. For instance, Lastauskaitė and Krušinskas (2021) pointed out that the increase in the adoption of digital technologies is directly linked to the improvement of productivity and economic activity in the manufacturing sector, as well as to a general growth of the Gross domestic product (GDP) in the EU (Kravchenko et al. 2019). Nonetheless, the impact of AI on youth employment and the debate on whether the shift of the job market towards more AI-centric roles may negatively influence young people is contested. For example, Aswathy et al. (2021) argue that AI will cause job displacement of low-skilled occupations due to the automation of routine tasks and the increase in the demands for technological competencies, thus directly affecting lower-income young people who cannot afford said qualifications. Whereas authors like Lu (2022), state that it is the skilled labor force who is seen affected negatively by the invention of AI, and vice versa. In other words, the arrival of AI presents a paradoxical scenario: while it offers remarkable advancements and efficiencies in productivity and economic growth, it simultaneously poses challenges for workforce integration, especially among the youth.

In terms of social issues of young people, digitalisation and AI have revolutionised their educational landscape. Again, AI is a double-sided

[1] The researchers would like to extend a special thanks to all the organizations that participated and made this research possible. On a voluntary basis, explicit mention (in alphabetical order) is made to: ASPAYM Castilla y León, Department of Youth Affairs of the Education and Youth Board of Estonia, Digital Child Rights Foundation, Eurocities, Eurodesk Spain, Extremadura Youth Council, and Spanish Youth Council.

sword in the sense that it can act as an enabler by facilitating and democratizing access to information, but it can also have a negative impact on the quality and reliability of the information that they consume, and it risks contributing to social isolation and mental health challenges such as depression and anxiety among youth (Grové 2021; Ştefan 2023a).

Although digitalisation has broadened access to information, and young people are the primary consumers of the internet —according to 2023 data, nearly 99% of individuals aged 16-24 in the EU reported using the internet daily, which is 4 percentage points higher than the usage rate among those over 25 (Eurostat 2023)— the benefits are not evenly distributed. Many sectors with limited resources struggle to afford these technologies (Ştefan 2023a). Therefore, AI may hinder social inclusion by perpetuating existing socio-economic divides, favoring those with access to technology and leaving behind those without (Park and Humphry 2019). Therefore, as we will explore in the following sections, youth work must include a balanced and informed use of AI in order to squeeze its potential to the fullest to mitigate the aforementioned inequalities and increase social inclusion.

Digital technologies have slowly taken over all aspects of life, and youth work is no exception (Pawluczuk and Șerban 2022). Youth workers have been integrating AI into their activities because its uses are many-fold (Șerban et al. 2020). The literature has identified many of the functions that AI can serve to the benefit of youth, especially young people with fewer opportunities, in the context of youth work and non-formal education. Given the rapid development of these technologies, this list is by no means exhaustive. Șerban et al. (2020) noted that AI-powered tools can:

— facilitate a deeper understanding of young people's needs and foster closer relationships;
— enable the early detection and mitigation of potential risks, such as stress indicators, thus enhancing mental health support;
— use specific digital tools such as heart rate monitoring through devices such as smart watches, which are particularly useful for people with dyslexia, anxiety, Down syndrome, autism and similar conditions;
— provide more customised advice and direction for young individuals' professional, social, and personal growth;
— augment educational systems to better align with the learning preferences and needs of youth by employing technologies

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

that track and analyse educational engagement and learning patterns;
— support young people with disabilities, such as applications that convert written text into speech for those with visual impairments, enhancing their accessibility to information and communication;
— AI-powered technology presents new possibilities by identifying, discouraging, and preventing online hate speech; and
— accelerate the delivery of youth services through the use of chatbots.

This paper will focus on the last point: the use of AI chatbots for digital youth services for digital youth information, advice, and counselling in Europe. In this highlight, we will present the development and implementation issues of AI chatbots. Also, the possible effects of deployment on the youth and the larger digital youth perspectives in Europe will be covered. The study will involve consultation of youth organizations in order to paint an all-encompassing picture about the effectiveness and prospects of digitalisation of youth services.

AI chatbots are increasingly being employed to facilitate interactive and integrated environments (Mageira et al. 2022). In the case of youth work, they can supplement the role of social workers, as they can be used to expand the services to the people who cannot access the services that depend on the presence of a human being (Ştefan 2023a). As pointed out by Pawluczuk (2023), these chatbots have also been used to provide information and to carry out other routinary activities like extracting, copying, and inserting data, or filling in forms, therefore allowing the youth workers to concentrate on more personalized and impactful interactions with all young people engaging in these services.

The latest studies (Pawluczuk 2023; Stefan 2024; Solyst et al. 2023; Vetrivel et al. 2024) have highlighted the fact that these electronic equipments should be designed for and by young people. They should do this by providing information, resources, and support in an engaging and friendly style that suits them best. For example, chatbots are particularly effective in settings like mental health support, as they can provide evidence-based coping strategies and resources via interactive and relatable conversations. This approach is possible because young people are adept at using technology and comfortable with digital interactions, making them more likely to engage with and benefit from such tools (Grové 2021).

Besides, Väänänen et al. (2020) stated that the so-called Civic Chatbots or CivicBots can support young people to be involved in civic affairs and engage with societal issues. These chatbots not only make participation in community affairs for young people easy and accessible but also help in fostering equality, making them adequate tools for motivating the youth to freely give their opinions and actively participate in their localities.

Overall, websites with integrated chatbots are facilitating a new way of how youth services relate to their audience. They provide the quick and customized communication imperative for attracting young people. Chatbots are accessible constantly and are equipped to answer questions about particular services or activities or to provide information: from clarifying the process of enrolling in a university (Atmauswan and Abdullahi 2022) to answering adolescents' inquiries on drugs, sex, and alcohol in an anonymous way. The convenience of this nonstop availability not only makes service delivery more efficient but also gives young people the power of immediate access to the information they may need (Crutzen et al. 2011).

This article highlights a significant issue in the design of chatbots: the lack of consideration for the profile, needs, and rights of end-users, that is, young people (Ştefan 2023a). Digital technologies such as AI services are not power-neutral, but rather designed, run, and controlled by profit-driven companies. These companies are a 'third party' that influences the relationship and interaction between young people and youth workers and services (Pawluczuk and Șerban 2022). Chatbots are powered by algorithms that analyse data, and this data, with the argument of providing better services, can be traced back to the user, thus jeopardising the data protection and anonymity of young people (Siurala 2020). On the technical side, there is a huge problem of disinformation among young people on the functioning of these tools and their potential risks, due to the opacity, complexity and private ownership of the analysis of data like algorithms and AI (Ştefan 2023a; Siurala 2020).

End-user data protection is not the only risk that deserves awareness when it comes to AI and chatbots in the context of youth work. Another issue is that AI chatbots have been found to be biased, due to the way their algorithms are designed (Șerban et al. 2020). The datasets that form the basis of chatbots are often fed with social biases, which can lead to discriminatory predictions between one target class and another, which can undermine the rights of people, especially those on the margins based on race, gender and social status (McQuillan and Salaj 2021; Ştefan 2023a). The misinformation on the

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

mechanisms used to provide content can lead to the incapacity to discern between real and fake outputs, dissemination of propaganda, filter bubbles that increase polarisation, and even radicalisation (Ştefan 2023a).

As mentioned earlier chatbots are often designed without involving young people in the process, and therefore overlooking their needs (Șerban et al. 2020). There are several ways to ensure the participation of youth. For example, one way is through the use of the People's Council. Citing McQuillan and Salaj (2021), these councils are bottom-up democratic assemblies, in which everyone has an equal say about the matter being decided. When speaking about AI, People's Councils become a way of collectively questioning the reasoning of the machine. These horizontal structures would incorporate both young people and youth workers to change the way these valuable tools are designed to meet the needs and respect the rights of young people. Another way is to carry out participatory action research to shape the outcomes. This method is characterised by the participation of the target population, to enable their influence in the decisions that will affect their lives (Chaudron and Di Gioia 2022). Following this line, this paper will include the perspective of youth organizations that offer chatbot services to the youths.

2. **Research objectives and methodology**

This study firstly conducted a review of various literature sources to get an overview of the development of information, advice and guidance services for young people in Europe, as well as the integration of artificial intelligence and chatbots in these digital youth services. Secondly, a qualitative methodological approach based on in-depth interviews was used. The aim was to collect data to find out in which activities chatbots are used, what functionalities they have and what benefits (or risks) they offer to youth information, counselling and guidance services.

Non-probability, purposive and convenience sampling was used to select a sample of organizations representing young people or whose main function is to provide youth services in EU countries (see Table 1). Between April and June 2024, 8 interviews were conducted virtually (via Teams) with heads of departments responsible for making decisions about integrating chatbots into their organizations or with youth workers responsible for teaching young people how to use these conversational digital assistants.

Table 1.
Characteristics of the youth organizations interviewed

Type of organization	Country	Level of action
Service Providers	Belgium	European level
Service Providers	Estonia	National level
Service Providers	Spain	Regional level
Service Providers	Spain	National level
Service Providers	The Netherlands	Worldwide level
Youth Representation	Spain	Regional level
Youth Representation	Belgium	European level
Youth Representation	Spain	National level

Source: Own elaboration.

The interview guide[2] began by laying the groundwork to find out if organizations knew what a chatbot was, if they had ever used one, and what experiences, both positive and negative, they had had. The questions then focused on what characteristics a chatbot should have within youth information, advice, and counselling services and what activities these tools are currently being used for. Afterwards, the questions focused on the opportunities offered, the risks involved and the challenges that the use of chatbots in youth services will bring.

A documentary and content analysis were then carried out (using MAXQDA software) to categorize and systematize the information collected. This made it possible to identify trends and commonalities between all the organizations interviewed, as well as with other research (to contrast the perspectives of this study with other sources). In this way, the interviews with these youth organizations allowed us to delve deeper into two main aspects. First, their perspectives on the role chatbots play or will play in youth services. Second, to find out what opportunities and risks they perceive chatbots to bring to young people.

Finally, it should be noted that all interviewees participated in this research on a voluntary basis and gave their consent for the

[2] The interview guide was developed based on the literature review to combine both the aspects of chatbots and artificial intelligence with youth services.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

information collected during the interviews to be used for this study (and for their quotations to be used anonymously). In this sense, the information obtained, collected and classified within the parameters of the applicable data protection legislation, was used for the subsequent analysis of the main results.

3. Evolution of youth information, advice, and counselling services in Europe

Due to the overwhelming amount of available information online, occasionally unreliable, youth information, advice, and counselling services are one of the cornerstones of the transition of young people to adulthood. They accompany them and defend their right to —trustworthy— information[3] (Sildnik and Simon 2020). Through the provision of accurate data, these services enable young people to make well-informed choices, thus, developing the critical thinking skills that are essential in today's complex information landscape. Furthermore, they assist in achieving social environment diversity and benefits. They make it possible for all young people, no matter their socioeconomic background, to be full members of society and therefore actively participate in decision-making activities (Reina et al. 2020).

Nowadays, the term 'youth information and counselling' is an umbrella term that includes a wide range of services and activities, such as informing, counselling, supporting, coaching, training, peer-to-peer, networking, or referral to specialized services (Sildnik and Simon 2020). The profession of youth information worker is now well organized (Frith et al. 2021), but these services have not always been the way we know them in the modern days. In this section, we will explore the evolution of youth services from their beginning until today. This section will also dig into what are the expected changes, opportunities, and challenges that youth information and counselling services are going to experience in the future, according to literature.

Léargas (2017), in the 1950s, specialized youth information services in Europe came into existence. These services were created in Finland with the opening of information centres for the young internal migrants coming from the countryside to the cities. The aim was to

[3] This is recognised in the Universal Declaration of Human Rights, in the Convention on the Rights of the Child, in the European Convention for the Protection of Human Rights and Fundamental Freedoms and in the Recommendations n. (90)7, CM/Rec(2010)8 and CM/Rec(2016)7 of the Council of Europe.

prepare these youths for the new challenges and complexities that they were facing. Later, in 1961, the very first 'walk-in' centre, called the Young People's Consultation Centre, was established in London. This changed the way in which young people could gain access to professional help without having to make an appointment.

'Open door' services became a trend during the 1960s. In cities like Ghent (Belgium) and Amsterdam (The Netherlands), new centers such as the Info Jeugd Centre for Youth Information and counselling and the Young People's Advice Center (Jongerenadviescentrum), respectively, were set up. These centers were different from the common method of youth work that was usually formal, bureaucratic and medical-psychiatric in nature. In contrast to the traditional facilities, these new centers created an atmosphere of acceptance, where the youth would be free to express themselves and seek assistance in a nonjudgmental way; hence, the feeling of belonging and support grew stronger.

It was during the 1970s when providing young people with information became more popular in Europe and integrated into the broader youth work practices in the majority of the countries. Recognizing the crucial role of youth information and counselling was finally done at the first European Conference of Ministers Responsible for Youth, which was held in Strasbourg in 1985. Thus, here, these services were emphasized as priorities of the future European-level policy cooperation. Consequently, the Council of Europe established the Committee of Experts on Youth Information in Europe in order to advance these plans (Léargas 2017).

These organizations kept evolving and specializing with time. A great milestone occurred in 1986 with the creation of the European Youth Information and Counselling Agency (ERYICA). This independent European non-governmental, non-profit association aims to safeguard the right of young people to information. They adopted the European Youth Information Charter in 1993, which served as the roadmap of youth work. This document has been revised several times to keep up with the rapid changes of the last 30 years (Reina et al. 2020). Other initiatives, that were implemented for the promotion of youth access to information and resources, were the European Youth Card (EYCA)[4]

[4] The EYCA is a discount card focused on the encouragement of mobility and active citizenship among youth across Europe. Providing a wide range of discounts to young people typically between 13 and 30 years of age, the card gives access to a range of transport, cultural, lodging and services at reduced rates in many European countries across the continent.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

(1987), Eurodesk[5] (1990), Euroguidance[6] (1992), the European Youth Portal[7] (2004), the Youth Guarantee programme[8] (2013), Youth Wiki[9] (2015), European Solidarity Corps[10] (2016) or the EU Youth Coordinator[11] (2018).

Looking at the future, youth information, advice, and counselling services in Europe now have the difficult task of providing solutions to the aforementioned problems, related to the unprecedented increase of technology in young people's lives, the wave of fake news and misinformation, the shift to an online mode of counselling –which includes AI and chatbots. Therefore, youth information services should work one step ahead, anticipating the needs of young people, establishing prevention measures, and mainstreaming youth information and counselling in diverse youth policies (Reina et al. 2020)

4. **The incorporation of artificial intelligence and chatbots in digital youth services**

4.1. *Artificial intelligence*

When speaking about AI, McQuillan and Salaj (2021) describe it as a sum of biased data, as opposed to the idea of a conscious superintelligent

[5] Eurodesk is an information network that aims at offering young people and those working with them timely, accurate, and relevant information in the realm of learning, training, and youth mobility. It provides information and counselling to support young people to participate in different programmes and campaigns across Europe.

[6] Euroguidance is a network of national resource and information centres for educational and employment sectors across Europe.

[7] The European Youth Portal provides information for young people on among others, opportunities, initiatives, employment or youth policies, as well as resources for organisations and policymakers.

[8] The Youth Guarantee Programme is an initiative launched by the European Union to guarantee that all young people under the age of 25 receive a good-quality offer of employment, continued education, apprenticeship or traineeship within four months of becoming unemployed or leaving formal education. For instance, in Castilla y León (Spain), a group of Youth Information Officers were responsible for informing and registering interested young people.

[9] The Youth Wiki is a database created by the European Comission that offers information on national policies and best practices on youth related issues.

[10] The European Solidarity Corps is a programme that aims to promote opportunities to volunteer or work on solidarity projects for young people.

[11] The EU Youth Coordinator is a position created by the European Commission as part of the EU Youth Strategy 2019-2027 for the management of youth policies of the EU. The aim is to ensure that young people's voices are represented in decision-making processes.

entity. Other scholars go a step further and affirm that AI is more an ideology rather than just algorithms, due to the implications in regards to categorization and language (Vesa and Tienari 2022).

The concept of AI was first used in the 1950s but it was not until 2017 when AI gained momentum at an international level, with the initiation of governance processes. These procedures included academia, public institutions, and civil society organizations to debate the potential risks and benefits that AI pose to general society (Ştefan 2023a). It is not clear when AI started to be used in the field of youth work, although international organizations (EU-Council of Europe Youth Partnership 2022) also refer to the last decade as the period when the trend of integrating AI in educational and social services grew. A big event in terms of the acknowledgment of the role of AI in youth work services was the 2018 Symposium 'Connecting the dots: Young people, social inclusion and digitalisation', held by the EU-Council of Europe Youth Partnership. In this sense, the European Commission defines AI as:

> Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions –with some degree of autonomy– to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or internet of things applications) (High-Level Expert Group on Artificial Intelligence 2019).

On the other hand, the Council of Europe (Leslie et al. 2021, 7) also provides a technical definition of this discipline:

> AI systems are algorithmic models that carry out cognitive or perceptual functions in the world that were previously reserved for thinking, judging, and reasoning human beings.

Nonetheless, youth work still lacks a standard definition of the correct approach to the use of AI (Pawluczuk 2023). The lack of a standard definition for chatbots may arise from their diverse applications, rapid technological evolution, and varied architectures. This can lead to user confusion, regulatory challenges, and ethical and privacy concerns. A common definition is very much needed as young people recognize that AI is and will continue to be part of their daily lives (Chaudron and Di Gioia 2022). Building on this increasing

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

integration of AI in youth-oriented services, one prominent application that has emerged is the use of chatbots.

4.2. *Chatbots*

Essentially, a chatbot is a conversational agent that uses natural language to dialogue with its user (in either text or speech form) with the aim of substituting traditional scrolling, swiping, or button click interfaces of data service providers (Council of Europe n.d.). Chatbots may exist solely as software or be integrated into physical social robots (Følstad and Brandtzæg 2017; Väänänen et al. 2020).

The origin of chatbots can be dated back to the 1960s (Dale 2016). These were simple, superficial systems unable to understand context and with limited interaction capabilities. Since then, the technological revolution and the resulting proliferation of computers and smartphones have driven the evolution of chatbots into more advanced applications that can be implemented for youth services even without technical skills (Verke 2020).

There are a myriad of youth services to which chatbots can be applied. They can be used not only to provide basic information about services and procedures, but also for cognitive behaviour therapy of young people (Fitzpatrick et al. 2017), for mental health support (Kretzschmar et al. 2019; Grové 2021), to provide adolescents with advice on topics such as sexual education (Crutzen et al. 2011), for legal counselling about young people's rights (Morgan et al. 2018), or to help them with their transition to college (Atmauswan and Abdullahi 2022).

According to Verke (2020), chatbots are creating numerous positive changes in various directions: Technologically, chatbots allow independence from traditional constraints such as time and place, and facilitate the streamlining of interactions due to their ease of use. They automate and carry out daily tasks and offer a range of technical solutions. Chatbots for the younger generation are highly customised and easily accessible with the provision of anonymity. Despite the fact that chatbots can be biased according to how their algorithmic systems learn (Feine et al. 2020; Kostenius et al. 2024), these digital assistants have the advantage that they can be called at any time, never get tired and provide constant support.

In the case of counselling, chatbots can be used to merge the existing services into one, making implicit knowledge explicit and offering guidance that can help to enrich the work of the counselor.

The counselors of the future, through automation, will be able to free up some of their valuable working hours and plan to improve the quality of their services and master new skills. Furthermore, chatbots can be of assistance in the creation of stronger counseling relationships, which in turn can reach diverse target groups (Verke 2020).

The problem is that these services have such complex systems that, even for their developers, they are considered black boxes. In other words, the decision-making mechanisms of these algorithms are frequently unknown (Ştefan 2023b). Therefore, if not even the creators of such systems can fully grasp the implications that chatbots may have in relation to data gathering and analysis, let alone the general public, especially if we look at vulnerable young people who have not been included in the design process. This issue raises privacy concerns in the sense that both young people and youth workers may want to share internal sensitive documents that include personal trackable data (Ştefan 2023a) and a sense of unease due to the unknown future development of this technology (Pawluczuk 2023).

As noted by Verke (2020), technologically –apart from the earlier mentioned problems with data protection and security, and ethical dilemmas– the lack of empathy and the human touch of facial expressions and gestures which can lead to a perceived coldness and impersonality, are limitations for chatbots. Along with that, chatbots may face the problem of automatically identifying subtle information requirements.

For the youth, text-based communication with a chatbot can give rise to a lack of understanding due to the absence of vocal tone and physical signs. Finding the right services via chatbots can be difficult, especially when it involves a case of individual and special needs. An additional issue is whether young users can tell when they are chatting with a bot or a human, which may influence their level of trust and the quality of the conversation. Moreover, in the arena of counselling, the speed of chatbots may be at the expense of the quality of services provided. The values of the employees who design the chatbots might inadvertently be reflected in the bot's responses that sometimes do not match the users' needs and expectations (Verke 2020). This is also because most of these conversational agents have an adult-centered design, which does not allow for the provision of trustworthy and safe systems that guarantee the fundamental rights of children, adolescents and young people (Escobar-Planas et al. 2022).

To date, the use of AI chatbots is not legally bound by any international body. Nonetheless, the EU is on the right track. The EU AI Act (European Parliament 2024), despite not explicitly including youth

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

services in its regulations, lays a basis that could create a framework to eventually provide protection to young people by classifying some AI systems as high-risk when they impact fundamental rights and are related to education, employment, and access to services. Complementing this initiative, other steps that the EU has taken to protect youth from the risks of these technologies include, but are not limited to the following. First, the report *Conclusions of the Council and of the Representatives of the Governments of the Member States Meeting within the Council on Digital Youth Work by the Council of the European Union* (2019). Second, the Digital Education Action Plan 2021-2027 (European Commission n.d.) to help to adapt the education systems to the digital era. Third, the EU Strategy on the Rights of the Child and the European Child Guarantee, which includes the section "Children's rights in the digital environment" to protect them against online risks. Fourth, or the *EU's General Data protection Regulation* (European Parliament and Council of the European Union 2016) that focuses on the collection and processing of personal data, including AI and chatbots. Fifth the *European Guidelines for Digital Youth Work*, which provides a framework for integrating digital tools and practices into youth work (Digital Youth Work 2019).

It is crucial that in the process of developing this academic and legislative body, stakeholders such as youth councils and organizations are taken into consideration. Not only to be heard but to be listened to, as stated in Article 12 of the UN Convention on the Rights of the Child: "I have the right to be listened to and taken seriously". That is why, the next sections will focus on the testimonies of these organizations in regards to the use, opportunities and challenges of chatbots in their services.

5. **The perspectives of youth organizations**

5.1. *What is a chatbot?*

From our analysis, the first and foremost finding that emerges is the difficulty of having a clear idea of what a chatbot really is, or even knowing how to recognise when one is being used on a recurring basis. All the interviewees, in addition to giving a different definition of what a chatbot was, highlighted several examples where young people and youth workers were unaware that they were interacting with a chatbot until they were made to realize that the tool with which they exchange information was, in fact, a chatbot. The above highlights the

need for privacy provisions and limitations to be transparent to users, and for them to be reminded at any time (Kretzschmar et al. 2019).

> It is difficult for everyone to keep up with the topic of AI and chatbots. It is becoming more and more pressing as a topic. But we're barely understanding it, we're barely understanding what it is. It's in so many technologies, that many times... for example, the term chatbot... many colleagues told me that ChatGPT is not a chatbot, we need to categorize what is AI and chatbot, what is the difference: are all chatbots AI, but all AI are chatbots? (Youth representation organization)

Along the same lines, the organizations interviewed indicated that they are still in the process of establishing a clear and convincing position on artificial intelligence (and, therefore, on chatbots)[12]. This is due to the complexity of understanding what it is and what it isn't, the whirlwind of digital tools being updated and the inability to keep up to date with all the issues related to artificial intelligence.

> We don't see AI and chatbot as a single area [...] it goes in the sense that this is evolving very fast, and administrations and services don't have the capacity to know everything, or to update [...] and to see that all levels and services have the capacity [to update] is still a challenge… (Service provider organization)

5.2. *From giving information about a youth centre to getting involved in mental health issues*

Interviewees noted a duality in the daily use of chatbots. Firstly, because they perceive that this type of tool is used by young people to solve administrative, logistical and informational queries within youth services, such as: finding out the opening hours of the youth centre, how to book a working room, what documents are needed to create a youth association, what cultural activities exist in their locality or how to participate in European volunteering.

> If I want to create a youth association, what do I have to do? If the room in the cultural centre is free and I want to book it, where

[12] It is considered necessary to point out that several organizations declined to participate in this research due to the fact that they do not yet have a position on AI and Chatbots.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

> do I have to book it? This seems ideal to me. [...] On mental health issues. Lately, there are a lot of self-help websites for young people with mental health problems (Youth representation organization).
>
> With tweezers, in emotional support, because in these issues, it is unpredictable that there is a person behind and not just a chatbot... Access to resources of all kinds, from an online shop, focused on young people, to find products… (Service provider organization).

Secondly, interviewees also perceived that the tools went from providing simple administrative information to directly addressing mental health issues. In other words, they stressed that there was no intermediate step in these tools and that many of them, which were not specifically designed for this purpose, ended up addressing problems of stress, anxiety or depression in children, adolescents and young people. Dosovitsky and Bunge (2023) conducted research with young people aged 13-18, testing a chatbot designed to psychoeducate young people about depression, teach behavioural activation and change negative thoughts. Although participants said that these conversational advisors could be positive for mental health, they highlighted technical and stylistic issues that developers should consider.

5.3. *Tailored for services and young people*

According to the interviewees, getting young people to engage with a chatbot depends not only on whether it is useful to them or whether it does not keep repeating the same answer in a loop. It is also important that the responses are adapted to both the service being offered and the language young people use. If a chatbot responds in the way a young person would interact, it may mean that they will either use the tool again or, on the contrary, discard it altogether.

> Adults or people that developed the systems are not young people. They do not know the slang they are using, or the specific topics that they need. Otherwise, they will not work… […] Otherwise, they will not use it anymore… This is the hardest part, to train and to create the information or the answers that are connected with what the young people are asking (Service provider organization).

In the same line, interviewees point out that chatbots have the capacity to facilitate access to youth services because they can be

consulted any day and at any time. Moreover, if their designs from the outset have an inclusive approach, they can also further enhance the accessibility of youth services to profiles that are often marginalised, especially if they respond in multiple languages. Due to the above, interviewees highlight the importance of young people and youth organizations being involved in co-design and testing processes of chatbots.

> There is a chatbot that gives you information in sign language... [The chatbot] has to be inclusive, with young people from different backgrounds, from other ethnicities, not to give a homogenous image of [the young people] who can access the programmes.... (Service provider organization).

6. **Opportunities and challenges of AI chatbots for digital youth services**

6.1. *Generator of ideas and organiser of time*

One of the opportunities offered by chatbots, according to interviewees, is the possibility of generating new ideas from a base. Being able to constantly share information with the chatbot, and have it responded, allows not only to generate ideas, but also to structure and systematise them in a clearer way. At the same time, they also believe that chatbots can be useful in taking on bureaucratic or time-consuming tasks, so that both young people and youth services can better manage their time.

> The ability to delegate processes to AIs... There are other issues that could be time-consuming or an administrative burden, so chatbots can fulfill bureaucratic issues. They can support a lot of grant applications, or reports... and bear the heavy burden of bureaucracy… (Youth representation organization)

Along the same lines, interviewees also emphasise that chatbots allow for the creation of organizational and working methodologies that are necessary for youth services to succeed. In other words, having a tool that allows you to plan activities, create gamified roadmaps or identify other technological tools, seems to be an opportunity that youth organizations and services should take into account.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

> Chatbots are being used for sure. In terms of non-formal education, it is especially useful in gamification of educational materials, or in simulation for policy practices, or for multiple prompting, the AI came with a lot of options and saved a lot of time (Youth representation organization).

Another of the opportunities offered by AI, according to interviewees, is the possibility of creating a model that generates various scenarios to anticipate what impacts the implementation of one service or another might have. In this sense, chatbots that can analyse data, generate scenarios and establish processes can improve the decisions that are made both within youth services and the actions of youth workers.

> If you have a model of [your services], before you close a street or demolish a building, you can simulate the scenario before you make that decision and find the best solution, before you put the money or dissatisfied the citizens.... (Service provider organization).

6.2. *Trust is earned through people and truthful information*

All interviewees agree that chatbots often fail because they do not give the right information or because it is noticeable that 'no person' has filtered what could be said (and what could not). This seems very relevant, especially when we are talking about such a heterogeneous group as young people. Especially because it is not the same how a child, an adolescent or a young person speaks and interacts. Therefore, one of the great challenges is to adapt the same information for several profiles and to be aware of the trends in each of them.

> The most important thing would be to delimit who can use this chatbot and who cannot [...] If this chatbot can be used by a person of legal age or if it can be used by children, adolescents... Based on this, I understand that the contents of the chatbot would have to be adapted to this public, in order to safeguard the integrity of all people… (Youth representation organization).

At the same time, interviewees point out that young people trust those chatbots, or the services they offer, when they know that a specialised person has been behind them doing a critical review of the information that the chatbot could or could not give. This trust also increases when young people know that an organization that already offers the same services face-to-face is behind the chatbot.

> The Child Help Line Chatbot, which is connected to child services, and young people trust it because they know that there are people behind it... Even if the answer is automatic, they know that there are people, specialists behind the service, and they trust in using it. They want to know who is behind the chatbot so they can trust and use it, instead of those big, general chatbots… (Service provider organization).

6.3. *No time, no resources, no competences*

Despite the above, interviewees also agree that creating a reliable chatbot is a present and future challenge. Firstly, because it involves a lot of human, technological and economic resources to keep a chatbot up to date both in the information it provides and in adapting from time to time to the reality of the young people themselves.

> We don't have enough human resources to create the chatbots... The specialists behind the chatbot might not be enough, they might not be prepared enough to design them [...] We are not prepared to coordinate the IT service… It can cost a lot of money to have a chatbot [...] We don't think about where the servers will be, how they will be protected… (Service provider organization).

In this sense, the interviewees underline that another challenge not only for chatbots but also for AI, is to keep up with all the technological advances that occur every second. According to Park and Lee (2024), it is necessary to improve the sustainability of chatbot services based on their artificial intelligence (in personalization and social aspects) and systemic factors (in their responsiveness and compatibility). The above is of paramount importance, as young people, youth workers and youth services do not seem to have enough time and resources to acquire all the skills they would need to critically use a chatbot. For that reason, organizations fear not so much that chatbots may provide wrong information, but that users do not have the time or the tools to check whether what they say is true, misleading or false.

> What characteristics would a person need to have in order to use it and benefit from it, in education and youth services? Is it necessary for young people to have critical-thinking and fact-checking skills in order not to misuse it? (Youth representation organization).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

Along these lines, interviewees also expressed concern about the side effects of such constant use of technology, especially when it is a tool that pretends to be a person. The opportunity provided by the immediacy of access to information may also mean that young people are increasingly unprepared to deal with situations where there is no immediate response. Likewise, youth organizations also seem to be concerned that this type of technology is increasingly leading to an impersonal tendency in social relations between young people themselves, as well as in their social and working relationships

> Applicants use it more and more for CV and Cover Letter. And not very skillful, I would say. And it is difficult for us, for recruitment, for writing production, it is very difficult to know who we are contacting… (Youth representation organization).

6.4. *Who is after our data?*

Interviewees agree that the biggest risk with chatbots is ensuring that our data is protected when interacting with these digital tools. Even when a chatbot is free in its functionalities, the way to generate revenue is often by selling user data. In this sense, interviewees also point out the difficulty in finding out who is behind certain chatbots and what their purposes are with our information. At the same time, it is also highlighted that young people may not be aware that the information they share is sensitive.

> The issue of profit, gain and development, poses issues in terms of why it is done and what people gain in this software […] It may not be in their interest to make (youth services) accessible to all. […] Discrimination bias and surveillance is very connected to who owns the monopoly and who has the data, who makes money out of it. It is a grey area (Youth representation organization).
>
> (The biggest risk is) The issue of data protection and privacy. Many companies want this information for their own profits. Young people are not aware of their private data and sell it easily. These chatbots can be very contracted, and their purposes can be different from what they seem. (Service provider organization).

In the same vein, interviewees also stress that the large companies that design these chatbots may have too much decision-making power. For example, determining which services are launched to the entire public or which are kept for a specific group; or which services

are made accessible and which are not. Likewise, these large corporations may be able to lobby to be the only ones to offer the technological solutions to carry out youth services in the digital world and not knowing whether the data they have has been ethically collected.

> The data has to be representative, because if we give decision making process and power to AI, the data has to be of good quality... [However] with AI the question is: the model is only good based on the data you have... If the data is good, has it been ethically sourced? (Service provider organization).

Finally, the organizations interviewed refer to the fact that the information that is shared and extracted from chatbots is not regulated. They therefore hope that, in view of the challenges they have been pointing out, the European Commission will regulate these tools and processes through the 'Artificial Intelligence Act'.

> It is true that chatbots, and the information that they generate, that we extract from these chatbots, is not regulated... So, it would be necessary for the European Commission to regulate it in this sense... The AI Act that is in the process of being drafted should contemplate this situation and the use of chatbots… (Youth representation organization).

Conclusions and Recommendations

All in all, the use of AI chatbots in the provision of digital information, advice and counselling services to the youth in Europe has several opportunities. Firstly, they have the ability to reach a larger number of users and provide young people with instant access to essential information. Secondly, they can contribute significantly to making youth services more inclusive and efficient (if they are designed with this in mind from the outset). Thirdly, they can help to reduce the administrative burden on youth workers and thus enable them to provide more responsive services. Fourthly, they can contribute to better decision-making in both the design and delivery of youth services, in order to mitigate negative outcomes and provide quality interventions.

On the other hand, there is no doubt that one of the main risks associated with AI chatbots is the lack of ethical principles in both their development and use within youth services. This is due to the fact that

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 127-154
https://doi.org/10.18543/djhr.3192 • http://djhr.revistas.deusto.es/

most are designed from an adult-centric approach, as well as the lack of mechanisms to ensure transparency, accountability and responsibility (Atkins et al. 2021). There is also the challenge of ensuring that conversational assistants do not become a substitute for human interaction and experience. In particular, a long-term implication is that young people may seek mental health support from chatbots first, rather than from trained professionals (Kooli 2023).

In addition to these risks, there is also the digital gap when using several services through a chatbot. Although young people are becoming increasingly connected, there are still a considerable number of young people who do not have access to the internet or a mobile device. According to Eurostat (2024), in 2019 only 52.35% of young people aged 16-29 had access to the internet at home via a computer and almost 10% did not have access via their mobile. Therefore, the incorporation of a chatbot has to be a complement and not a replacement of a service, in order to ensure that all people can access services both online and offline.

Building on the points above, here are some recommendations to consider when developing and using AI chatbots to benefit young people. Policymakers should ensure sufficient regulations are in place regarding the use of AI chatbots, data protection, and ethical standards, especially concerning youth data, so that this information is not sold to the highest bidder or used to monitor users (Pawluczuk 2023). Additionally, as mentioned, training programmes for young people and youth workers focused on digital literacy and critical thinking would be valuable to promote the responsible use of these technologies (Digital Youth Work 2019). Youth organizations can play a crucial role by involving young people in chatbot development stages, aligning tools with their needs, and conducting research on the impact of AI across diverse youth contexts and age groups to highlight the varied needs of young users (Dosovitsky and Bunge 2023). Practitioners should regularly update chatbots based on feedback from youth organizations to ensure cultural sensitivity, inclusivity in language and content, and build trust by incorporating human oversight of chatbot processes. It is important to ensure that chatbots are designed from an inclusive and intersectoral approach to avoid discrimination and biased decision-making that excludes any profile of young people (Väänänen et al. 2020).

On the other hand, the present research also has some limitations. The first is the impossibility to generalise the findings obtained to the reality of the EU and youth services in this region, due to the small sample size of the current study. In particular, due to the impossibility

of interviewing one organization per member country and per type of youth service. The second, similar to the previous one, was not being able to count with young people during this study, in order to know in depth their perspectives, expectations and concerns about these issues. Therefore, it is recommended that future research overcome these limitations to continue generating evidence that will help the design of chatbots within youth services.

In conclusion, the current study shows that chatbots can be a positive tool within youth services, especially if they are co-designed with youth organizations and young people to minimize threats and maximize their effectiveness. However, it seems imperative to consider and address potential risks when developing AI chatbots, especially if they are to become effective tools for expanding coverage, improving productivity and increasing the effectiveness of youth services. In other words, it is crucial to ensure ongoing evaluation and monitoring of AI chatbots in youth services to ensure that they achieve their intended goals and do not cause unintended harm.

References

Aswathy, Karanath A., Rosalind Gonzaga, and Josna S. Francis. 2021. «A study on the awareness of artificial intelligence among youth and its impact on employment». *International Journal of Advanced Research in Science, Communication and Technology* 5 (1): 461–465. doi.org/10.48175/IJARSCT-1168.

Atkins, Suzanne, Ishwar Badrie, and Sieuwert van Otterloo. 2021. «Applying ethical AI frameworks in practice: evaluating conversational AI chatbot solutions». *Computers and Society Research Journal 1.*

Atmauswan, Pica S., and Akibu M. Abdullahi. 2022. «Intelligent chatbot for University Information System Using Natural Language Approach.» *Advanced Science and Business Journal* 3 (2): 59-64.

Chaudron, Stephane, and Rosanna Di Gioia. 2022. *Artificial ntelligence and the rights of the child. Young people's views and perspectives.* Luxembourg: Publications Office of the European Union.

Council of Europe. n.d. *Glossary: Artificial Intelligence (AI)*. Council of Europe. Accessed May 13, 2024. https://www.coe.int/en/web/artificial-intelligence/glossary#:~:text=ARTIFICIAL%20INTELLIGENCE%20(AI),abilities%20of%20a%20human%20being.

Council of the European Union. 2019. «Conclusions of the Council and of the Representatives of the Governments of the Member States Meeting within the Council on Digital Youth Work.» Official Journal of the European Union, C 414: 2. December 10. Accessed April 20th, 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019XG1210(01).

Crutzen, Rik, Gjalt-Jorn Y Peters, Sarah Dias Portugal, Erwin M. Fisser, and Jorne J. Grolleman. 2011. «An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study.» *Journal of Adolescent Health* 48 (5): 514–519.

Dale, Robert. 2016. «The return of the chatbots.» *Natural Language Engineering* 22 (5): 811–817.

Digital Youth Work. 2019. *European Guidelines for Digital Youth Work*. YouthLink Scotland. Accessed November 6, 2024. https://digitalyouthwork.eu/guidelines/.

Dosovitsky, Gilly, and Eduardo Bunge. 2023. «Development of a chatbot for depression: adolescent perceptions and recommendations.» *Child and adolescent mental health* 28 (1): 124-127.

Escamilla, Alonso, and Irina Lonean. 2021. *Briefing 4: Review of research on the impact of COVID-19 on youth work, youth organisations, and the digitalisation of services and activities for young people.* Strasbourg: Council of Europe.

Escobar-Planas, Mariana, Emilia Gómez, and Carlos. D. Martínez-Hinarejos. 2022. «Guidelines to develop trustworthy conversational agents for children.» *arXiv preprint* arXiv:2209.02403.

EU-Council of Europe Youth Partnership. 2022. *Research on AI and Young People.* EU-Council of Europe Youth Partnership. Accessed May 2, 2024. https://pjp-eu.coe.int/en/web/youth-partnership/research-on-ai-and-young-people.

European Commission. n.d. *Digital Education Action Plan (2021-2027)*. European Education Area. Accessed November 6, 2024. https://education.ec.europa.eu/es/focus-topics/digital-education/action-plan.

European Parliament. 2024. *P9 TA(2024)0138. 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.

European Parliament and Council of the European Union. 2016. *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 (General Data Protection Regulation)*. Official Journal of the European Union, L 119: 1-88. Accessed April 20th, 2024. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Eurostat. 2023. *96% of young people in the EU use the internet daily*. Accessed May 12, 2024. https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20230714-1#:~:text=In%202022%2C%2096%25%20of%20young,94%25%20in%20all%20EU%20countries.

Eurostat. 2024. *Individuals. Mobile internet sccess*. Accessed May 24, 2024. https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_im_i__custom_11559915/default/table?lang=en.

Feine, Jasper, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2020. «Gender bias in chatbot design.» In *Chatbot research and design third international workshop, CONVERSATIONS 2019*, Amsterdam: The Netherlands, November 19–20, 2019, Revised Selected Papers, 79-93. doi.org/10.1007/978-3-030-39540-7_6

Fitzpatrick, Kathleen K., Alison Darcy, and Molly Vierhile. 2017. «Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial.» *JMIR Mental Health* 4 (2): e19.

Følstad, Asbjørn, and Petter Bae Brandtzæg. 2017. «Chatbots and the new world of HCI.» *Interactions* 24 (4): 38–42. doi.org/10.1145/3085558.

Frith, Audry, Eva Reina, Imre Simon, and Safi Sabuni. 2021. *YouthInfoComp: Youth information worker competence framework.* Luxembourg: ERYICA and Eurodesk.

Grové, Christine. 2021. «Co-developing a mental health and wellbeing chatbot with and for young people.» *Frontiers in Psychiatry* 11: 606041.

High-Level Expert Group on Artificial Intelligence. 2019. *A definition of AI: Main capabilities and scientific disciplines.* Brussels: European Commission.

Kooli, Chokri. 2023. «Chatbots in education and research: A critical examination of ethical implications and solutions.» *Sustainability* 15 (7): 5614.

Kostenius, Catrine, Frida Lindstrom, Courtney Potts, and Niklas Pekkari. 2024. «Young peoples' reflections about using a chatbot to promote their mental wellbeing in northern periphery areas-a qualitative study.» *International Journal of Circumpolar Health* 83 (1): 2369349. doi.org/10.1080/22423982.2024.2369349

Kravchenko, Olena, Maryna Leshchenko, Dariia Marushchak, Yuriy Vdovychenko, and Svitlana Boguslavska. 2019. «The digitalization as a global trend and growth factor of the modern economy.» *SHS Web of Conferences* 65: 7004. doi.org/10.1051/shsconf/20196507004.

Kretzschmar, Kira, Holly Tyroll, Gabriella Pavarini, Arianna Manzini, and Ilina Singh. 2019. «Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support.» *Biomedical Informatics Insights* 11: 1-9. doi.org/10.1177/1178222619829083

Lastauskaite, Aiste, and Rytis Krusinskas. 2021. «Impact of digitalization factors on EU economic growth.» In *2021 IEEE International Conference on Technology and Entrepreneurship (ICTE)*, 1–6. Kaunas: IEEE. doi.org/10.1109/ICTE51655.2021.9584695.

Léargas, Per. 2017. «A history of youth information in Europe.» *Léargas Blog*. Accessed May 14, 2024. https://blog.leargas.ie/blog/a-history-of-youth-information-in-europe.

Leslie, David, Christopher Burr, Mhairi Aitken, Josh Cowls, Mike Katell, and Morgan Briggs. 2021. *Artificial intelligence, human rights, democracy, and the rule of law: a primer*. Strasbourg: Council of Europe.

Lu, Chia-Hui. 2022. «Artificial intelligence and human jobs.» *Macroeconomic Dynamics* 26 (5): 1162–1201. doi.org/10.1017/S1365100520000528.

Mageira, Kleopatra, Dimitra Pittou, Andreas Papasalouros, Konstantinos Kotis, Paraskevi Zangogianni, and Athanasios Daradoumis. 2022. «Educational AI chatbots for content and language integrated learning.» *Applied Sciences* 12 (7): 3239. doi.org/10.3390/app12073239

McQuillan, Dan, and Reka Salaj. 2021. «Precarious youth and the spectre of algorithmic stereotyping.» In *Young people, social inclusion and digitalisation: Emerging knowledge for practice and policy*, edited by Moxon et al., 87-103. Strasbourg: EU-Council of Europe Youth Partnership.

Morgan, Jay Paul, Adeline Paiement, Jane Williams, Adam Zachary Wyner, and Monika Seisenberger. 2018. «A chatbot framework for the children's legal centre.» In *Legal knowledge and information systems*, edited by Monica Palmirani, 205–209. Amsterdam: IOS Press.

Park, Arum, and Sae Bom Lee. 2024. «Examining AI and systemic factors for improved chatbot sustainability.» *Journal of Computer Information Systems* 64 (6): 728-742. doi.org/10.1080/08874417.2023.2251416

Park, Sora, and Justine Humphry. 2019. «Exclusion by design: intersections of social, digital and data exclusion.» *Information, Communication & Society* 22 (7): 934–953. doi.org/10.1080/1369118X.2019.1606266.

Pawluczuk, Alicja. 2023. *Automating youth work: youth workers views on AI*. Strasbourg: EU–Council of Europe Youth Partnership

Pawluczuk, Alicja, and Adina Mariana Șerban. 2022. *Technology and the new power dynamics: Limitations of digital youth work.* Strasbourg: EU-Council of Europe Youth Partnership and Council of Europe's Youth Department.

Reina, Eva, Audry Frith, Safi Sabuni, and Imre Simon. 2020. *Greening youth information services: A guide developed by ERYICA and Eurodesk.* Luxembourg: ERYICA and Eurodesk.

Șerban, Adina M., Dan Moxon, Dunja Potocnik, Lana Pasic and Veronica Ştefan. 2020. «An overview of social inclusion, digitalisation and young people.» In *Young people, social inclusion and digitalisation: Emerging knowledge for practice and policy*, edited by Moxon et al., 87-103. Strasbourg: EU-Council of Europe Youth Partnership.

Sildnik, Hannes, and Imre Simon. 2020. *Youth information and counselling in Europe in 2020.* Luxembourg: ERYICA.

Siurala, Lasse. 2020. *Youth work and techlash: What are the new challenges of digitalisation for young people?* Strasbourg: EU-Council of Europe Youth Partnership.

Solyst, Jaemarie, Ellia Yang, Shixian Xie, Amy Ogan, Jessica Hammer, and Motahhare Eslami. 2023. «The potential of diverse youth as stakeholders in identifying and mitigating algorithmic bias for a future of fairer AI.» *Proceedings of the ACM on Human-Computer Interaction* 7, no. CSCW2: 364: 2 - 364: 27.

Ştefan, Veronica. 2023a. *Shapers & shakers: Young people's voices in the world of artificial intelligence.* Strasbourg: EU-Council of Europe Youth Partnership.

Ştefan, Veronica. 2023b. «Digitalisation and AI in the youth sector: Reality or hype?» *Coyote Magazine*, Issue 36. Strasbourg: EU-Council of Europe Youth Partnership.

Stefan, Veronica. 2024. *Insights into artificial intelligence and its impact on the youth sector*. Strasbourg: EU–Council of Europe Youth Partnership

Väänänen, Kaisa, Aleksi Hiltunen, Jari Varsaluoma, and Iikka Pietilä. 2020. «CivicBots – Chatbots for supporting youth in societal participation.» In *Chatbot research and design*, edited by Asbjørn Følstad, 143–157. Cham: Springer. doi.org/10.1007/978-3-030-39540-7_10

Verke. 2020. *What the bot? Chatbots in youth work*. Accessed May 2, 2024. https://www.verke.org/en/blogs/what-the-bot-chatbots-in-youth-work/.

Vesa, Mikko, and Janne Tienari. 2022. «Artificial intelligence and rationalized unaccountability: Ideology of the elites?» *Organization* 29 (6): 1133-1145. doi.org/10.1177/1350508420963872

Vetrivel, Sandhya, Chandrasekar Sowmiya, Patel Arun, Pandi Saravanan, and Rangasamy Maheswari. 2024. «Guiding principles for youth-centric development: Ethical AI.» In *Exploring youth studies in the age of AI*, edited by Zeinab Zaremohzzabieh, Rusli Abdullah and Seyedali Ahrari, 298-314. Hershey: IGI Global.

The human right to participate and its connection to artificial intelligence

El derecho humano a participar y su conexión con la inteligencia artificial

María Dolores Montero Caro
Universidad de Córdoba. España
mdmontero@uco.es
Orcid: https://orcid.org/0000-0001-9033-620X

https://doi.org/10.18543/djhr.3193

Submission date: 30.05.2024
Approval date: 02.12.2024
E-published: December 2024

Cómo citar / Citation: Montero, María Dolores. 2024. «The human right to participate and its connection to artificial intelligence.» *Deusto Journal of Human Rights*, n. 14: 155-172. https://doi.org/10.18543/djhr.3193

Summary: Introduction. 1. The right to participate as a human right. 2. Linking the development of technology to political participation in democracy. 3. A future perspective on artificial intelligence and its democratic impact. Conclusions. References.

Abstract: This article analyses the right to participate in democracy as a human right and its link to the development and implementation of artificial intelligence. First, it explores the fundamental aspect of the right to participate as a human right in the democratic framework, reflecting on its importance and its basic function of generating open spaces for the debate and presentation of other rights, highlighting that political participation generates a pull effect on other rights as citizens are in a favourable position for the defence and recognition of their rights. Next, emphasis is placed on the role of emerging technologies in facilitating and enhancing democratic engagement, bearing in mind that technological development has a direct influence on all aspects of people's lives, so that democracy in general, and each society's methods of organisation in particular, are also affected. Finally, a significant part of the discussion revolves around the future perspective of artificial intelligence and its potential impact on democracy, exploring both favourable developments and potential challenges. Artificial intelligence will undoubtedly continue to conquer more spheres of human endeavour, so it is worth reflecting on the importance of adapting this new technology to the future of democracies, while respecting its essential elements and guaranteeing citizens' fundamental rights. Finally, the article concludes by summarising the main

ideas and implications, underlining the critical importance of safeguarding democratic principles in the midst of technological advances.

Keywords: participation, artificial intelligence, human rights, democracy, politics.

Resumen: Este artículo analiza el derecho a participar en democracia como derecho humano y su vinculación con el desarrollo e implementación de la inteligencia artificial. En primer lugar, se profundiza en el aspecto fundamental del derecho a participar como derecho humano en el marco democrático, reflexionando sobre la importancia del mismo y su función básica de generar espacios abiertos al debate y presentación de otros derechos, destacando que la participación política genera un efecto de atracción sobre otros derechos al estar la ciudadanía en posición favorable para la defensa y reconocimiento de sus derechos. A continuación, se hace hincapié en el papel de las tecnologías emergentes a la hora de facilitar y mejorar el compromiso democrático, teniendo en cuenta que el desarrollo tecnológico tiene una influencia directa en todos los órdenes de la vida de las personas, de manera que la democracia en general, y los métodos de organización de cada sociedad en particular, se ven también afectados. Por último, una parte significativa de la discusión gira en torno a la perspectiva futura de la inteligencia artificial y su impacto potencial en la democracia, explorando tanto los avances favorables como los desafíos potenciales. Sin duda la inteligencia artificial va a seguir conquistando más esferas de actuación humanas por lo que cabe reflexionar sobre la importancia de adaptar esta nueva tecnología al porvenir de las democracias, respetando sus elementos esenciales y garantizando los derechos fundamentales de los ciudadanos. Por último, el artículo concluye resumiendo las principales ideas e implicaciones, subrayando la importancia crítica de salvaguardar los principios democráticos en medio de los avances tecnológicos.

Palabras clave: participación, inteligencia artificial, derechos humanos, democracia, política.

Introduction[1]

The fundamental right to political participation, in addition to being a human right enshrined in Article 21 of the Universal Declaration of Human Rights of 1948, shows that we are dealing with a basic right that belongs to every human being by virtue of the fact that they are human beings, and which, in turn, favours the possibility of generating democratic structures wherever there is a society that can organise itself as such (Castellanos 2020). It is a right that has been widely studied, both in doctrinal and jurisprudential terms (Ruiz Robledo 2018), which highlights its importance and relevance.

This is an irreplaceable element of any democratic scenario. Citizens must participate in public affairs and, as a consequence, democracy is permanently reasserting itself. Any hint of shadow or suspicion about this basic principle, any element that disturbs the general awareness that citizens participate in public affairs, has dire consequences. Recall President Trump's statements describing the election in which President Biden defeated him as "the greatest election fraud in history", provoking a riot and subsequent violent occupation of the Capitol that led to the loss of four lives and dozens of injuries. If just a few statements by the losing candidate could have led to those tragic incidents of January 2021, what could not happen if, on the occasion of introducing technological mechanisms in the development of the elections, the electronic voting could have been manipulated by the influence of some technological error or the incidence of artificial intelligence in defining the results (Castellanos 2024, 276-277).

It is therefore essential to reflect on how the basic foundations of democracy, citizen participation in the democratic arena, are affected by the technological elements that are gradually being introduced into the day-to-day life of citizens. Thus, although it is possible to state that the introduction of Artificial Intelligence (AI) in the democratic sphere

[1] This work has been carried out within the framework of the R&D&I Project PID2022-136439OB-I00/ MCIN/AEI/10.13039/501100011033, Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas, funded by the Ministry of Science and Innovation, Co-funded by European Regional Development Fund "A way to make Europe". And in the framework of the project of the Ministry of Science and Innovation (MICINN), Proyectos de Generación de Conocimiento 2022 (PID2022-140415NB-I00) "De la transparencia al Gobierno Abierto" (From Transparency to Open Government). This work is also part of the research activity of the SEJ-372 Research Group of the Andalusian Regional Government "Democracia, Pluralismo y Ciudadanía", of which the author is a member.

can have both positive and negative aspects, it is important to consider both sides of the coin. On the one hand, disruptive technologies, and specifically AI, could help improve the efficiency and transparency of democratic processes by using technologies to analyse large datasets to identify patterns and trends in voter preferences, as well as by employing *chatbots* and voice recognition systems to facilitate access to information and public services. However, there is also the potential for the incorporation of AI in the democratic sphere to raise concerns about confidentiality, protection and manipulation of information. For example, the use of AI technologies to collect and analyse personal data could compromise the privacy of individuals, while the influence of AI on political decisions could undermine the integrity and fairness of democratic processes.

Hence, the main reflection that we will address in this article should be about an adaptation of artificial intelligence that is harmonious with human reality. Because technological progress is as uncontroversial as it is debatable whether the insertion of artificial intelligence in all areas of life does not entail an associated risk.

1. **The right to participate as a human right**

Citizen participation in the political, social and economic affairs of a society is fundamental to the full exercise of democracy and respect for human rights. This principle, enshrined in numerous international human rights instruments, establishes that all individuals have an inherent right to participate in making decisions that affect their lives and communities. In this context, the right to participate stands as one of the fundamental pillars of a just and equitable society.

As we have noted, the Universal Declaration of Human Rights, adopted by the United Nations General Assembly in 1948, proclaims in article 21 the right of everyone to take part in the government of his or her country either directly or through freely chosen representatives. This right is complemented by other international instruments, such as the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights, which recognise and protect participation as an essential component of human dignity and the realisation of other fundamental rights (Viciano and González 2014).

Citizen participation manifests itself in a variety of ways, ranging from voting in democratic elections to participating in peaceful protests, expressing opinions in the media and social networks, and

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 155-172
https://doi.org/10.18543/djhr.3193 • http://djhr.revistas.deusto.es/

collaborating in civil society organisations. These practices not only strengthen the legitimacy of democratic institutions, but also empower individuals and foster a sense of belonging and responsibility towards the community.

However, despite the recognised importance of participation as a human right, there are numerous obstacles that can hinder its full exercise. These include discrimination, exclusion, political repression and lack of access to the information and resources necessary to participate meaningfully in public life. It is therefore the responsibility of states and the international community to ensure that the participatory rights of all individuals, especially historically marginalised or vulnerable groups, are respected and protected. Algorithms are created by people, who, either intentionally or unconsciously, may incorporate biases and stereotypes they hold about certain social groups into the systems they develop (Soriano 2021, 92). Moreover, to all these traditional elements affecting citizens' participation in public affairs must be added those associated with certain technological developments, notably the irruption of artificial intelligence in all spheres of people's lives, including, of course, in their political dimension.

It is undeniable that citizen participation is a fundamental pillar of any democratic political system. In democracy, crucial decisions affecting life in society are taken by citizens, either directly or through representatives. However, between these two options, the latter seems to be preferable. The representative democracy model, in which citizens elect a small group of representatives to debate and vote on behalf of the population, is more viable and effective than the direct democracy model, in which all decisions are subject to the constant scrutiny of the citizenry. Of course, the latter system would be practically unfeasible in large communities. And this is corroborated by the fact that this is an idea that is well established in the democratic imaginary of societies, but that emerging technologies are leading to the elimination of many physical and administrative barriers has brought it back into the debate (Sánchez 2006). Why not use technology to establish a daily plebiscite? Why not use technology to advance the democratic conception of societies? Of course, the use of technology could lead to democratic improvement in terms of expanding the possibilities for direct citizen participation (Ackerley 2017; López Rubio 2023; Álvarez and de Montalvo 2011).

In any case, in addition to practical issues, there are other substantive reasons for representative democracy, as this model encourages political decisions to be made in the general interest, rather

than for particular interests. Representatives are obliged to publicly justify their positions, which contributes to safeguarding the common good. In this context, electoral participation emerges as the main channel through which citizens intervene in public affairs, so that the election of representatives is the cornerstone of legitimisation of the political decisions they take. In this regard, it is important to highlight this premise in order not to lose sight of the fact that other mechanisms of participation should be seen as complementary to, but never as substitutes for, participation in the electoral process.

Moreover, the recognition of the right to participate in public affairs, generally conveyed through free elections in open and transparent democratic processes, implies the democratic consecration of any self-respecting society. We underline this point because when we deal with technological insertion in the democratic space and, therefore, its influence on citizens' political participation, we will be dealing with the integration of technological novelties in the public arena or of human incidence by definition. In this sense, the public space and citizen interrelation to discuss and debate public affairs is the place par excellence where the conception of man as a citizen, as an integral element of society, is embodied. This is not just a technological improvement, which is important in terms of improving the quality of life, as there are many, but rather it has an impact on the most essential part of the reality of citizenship. In terms of participation and recognition of rights, not only in their political conception, but also as a human right, the degree of democratic progress of any society is measured and founded. This is the reason why we advocate a serious and profound reflection in this work by not frivolously observing the possibility that artificial intelligence in particular, and disruptive technologies in general, may distort the very democratic conception that recognises the right to participate as a human right.

In short, the right to participate is an essential component of democracy and human rights, ensuring that all people could contribute to the development of their societies and to influence the decisions that affect their lives. Promoting and protecting this right is fundamental to building more just, inclusive and democratic societies. In this regard, the United Nations General Assembly has expressed itself through the approval in 2015 of the so-called 2030 Agenda for Sustainable Development, by virtue of which the Member States committed themselves to the achievement of 17 goals (Sustainable Development Goals, SDGs), including SDG16 "Peace, justice and strong institutions". This goal includes among its targets to ensure responsive, inclusive, participatory and representative decision-making

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 155-172
https://doi.org/10.18543/djhr.3193 • http://djhr.revistas.deusto.es/

at all levels (SDG 16.7), as well as to ensure public access to information and protect fundamental freedoms, in accordance with national laws and international agreements (SDG 16.10). As the SDG slogan "leave no one behind" states, opening up the world also means opening up decision-making to citizens, generating more participatory societies and more open institutions (Montero 2022).

2. **Linking the development of technology to political participation in democracy**

We have noted above that the human right to participate is, for the most part, realised through electoral processes. In this regard, we note that, in 2024, we are witnessing a historic milestone: more than 60 countries around the world hold national elections[2], making it the largest election year in history. However, this massive election period is marked by growing concerns about the role of social media and artificial intelligence in shaping public debate and democratic integrity. In particular, platforms such as *Meta*, with its 3 billion users on *WhatsApp*, *Instagram* and *Facebook*, are at the centre of attention because of their power to shape the information ecosystem globally.

In this sense, concerns about election-related disinformation and violence have led activists, legislators and journalists to sound the alarm bells (Carrillo 2022). This is not a fictitious concern; there is indeed a risk of undue influence on democratic processes through the pernicious use of social media. However, many platforms are changing their teams responsible for maintaining social media security, raising questions about their ability to address these issues effectively.

There have been previous examples that have shown that the influence of technology on democratic processes is, at the very least, something to reflect on. For example, in 2016, *Meta* became the focus of attention for its alleged role in the rise of Donald Trump to the presidency of the United States. And while the company invested in new tools and processes to combat disinformation during the 2020 presidential election, subsequent reports revealed that groups such as "Stop the Steal" continued to grow in the weeks following the election, questioning the legitimacy of Joe Biden's victory. This lack of

[2] In addition to the European Parliament elections scheduled for 9 June, many countries will be called to the polls in 2024, accounting for more than half of the world's population. These countries include, for example, the United States, Indonesia, Pakistan, Brazil, Russia and Mexico, among others.

effective action by *Meta* has been criticised, especially after events such as the Capitol Hill riot of 6 January 2021. And this is not an isolated case since, in Brazil, the attack by Bolsonaro supporters on government buildings on 8 January 2023 was also fuelled by the communication of audios and disinformation via *WhatsApp* and *Telegram*. As we can see, "disinformation can have serious consequences for democracies and national and international security, and in today's digital age the main place where it originates and spreads is in the virtual environment" (Montero 2023, 69). It is therefore not surprising that both the EU and national governments, including Spain, have focused in recent years on combating disinformation, with the approval of the Digital Services Act (DSA) in the EU and the Spanish government's National Security Strategy 2021, for example, standing out.

Also relevant are the case of the Philippines, where *Meta* facilitated the rise of former president Rodrigo Duterte by allowing *trolls* attacking the opposition to flourish (Ragragio 2021), and the case of Sri Lanka, where the platform allowed the promotion of content marked as false and inciting violence. These examples illustrate the global challenges faced by social media in protecting democratic integrity.

Therefore, we want to put on the table for debate the importance that the enshrinement of technological influence on voters' decisions can have in electoral processes. The case of Slovakia, with the manipulation of audios through artificial intelligence that could affect the voting in 2023, has already put many democracies in Europe on notice, so the control of technological cleanliness in the electoral scenario will be a key point for the maintenance of confidence in electoral processes.

In any case, the electoral manipulation of voters or even the possibility of generating situations in which the veracity of the results is in doubt are important problems, but not the only ones if we take into account the weight of new technologies in democratic spaces. Thus, other relevant problems in this field are determined by the tribalisation of politics in social networks. By political tribalisation in social networks we refer to the phenomenon in which users group together in virtual communities based on their ideological affinities, creating echo chambers (Pariser 2011) where opinions are constantly reinforced. It is quite evident that this can lead to polarisation and the rejection of opposing views, making dialogue and the search for common solutions difficult. Moreover, it is paradoxical that the confluence of increased channels of communication between human beings does not naturally bring diverse positions closer together, but rather clusters similar conceptions and erects impenetrable barriers to other ideas.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 155-172
https://doi.org/10.18543/djhr.3193 • http://djhr.revistas.deusto.es/

Election campaigns have become real information wars whose main stage is the social networks. In these battles, the victory does not lie with the best arguments, but with the most intelligent algorithms. This information war, which even encourages the spread of fake news with the aim of convincing the electorate to vote for a particular political option, is what Han has called 'infocracy' (Han 2022, 25).

This situation has been brought about by the development of emerging technologies, particularly because of the rise of social networks (Balaguer 2023). These tools, used appropriately, imply a better and greater encounter between citizens and their ideas, but, unfortunately, this is not only the case in this direction. Hence, we highlight tribalisation as an element to counteract the advance of technology and its political influence. Thus, arguing against this trend involves highlighting the importance of intercultural dialogue and empathy in overcoming the divisions created by tribalisation. These are elements to work on because the ideal democratic scenario is one in which a collective space is generated in which all contributions have a place and citizens are not excluded on the basis of their ideas. In particular, social networks have the technical potential to be able to establish this in our democracies, which is why it will be necessary to establish guidelines for action to avoid hatred and confrontation through the networks. Otherwise the distance between those who think differently will grow, as will this phenomenon of tribalisation.

Another noteworthy aspect relates to the multiplication of opinions and its link to what could be described as an intense democracy. Thus, while the multiplication of opinions can be seen as a positive aspect of digital democracy, where more voices have the opportunity to express themselves, this does not automatically guarantee a more intense or participatory democracy. In many cases, the overabundance of information can lead to confusion and scepticism, especially if the quality of opinions is low or if misinformation is spread. Here it is worth reflecting on a controversial aspect: whether to encourage the presence of all opinions on an equal footing or to consider that, while no one can be silenced in a democratic setting, not all opinions should be on an equal footing. Opinions based on prejudice, disqualifications, a complete lack of rigour or even outright lies should not be accorded the same level of respect as those freely expressed by other citizens. It is complicated to fit in because, unconsciously, silencing opinions implies a certain anti-democratic game. Hence, democratic pedagogy itself implies visualising that the necessity of participation implies, by its very nature, the responsibility of participation.

What we believe is that responsible citizen participation should be advocated, which is based on processes of information and the broadening of knowledge. Participation is not the mere externalisation of a conjecture without the parameters of reflection. Nor is it a debate among experts. It is a scenario in which citizens can freely participate and express their opinions. However, these opinions must be based on prior conceptions or knowledge that invite debate and collective reflection. The mere articulation of unreflective and unargued ideas does not have the same consistency and, consequently, does not nurture democracy or generate fruitful participation. Participation for the sake of participation does not lead to better democracy. Responsible participation does. Hence, the intensity of democracy does not go hand in hand with the grouping of opinions. Moreover, the ideal scenario, within representative democracy, is not so much participatory democracy understood as a sum of numbers, but deliberative democracy, as a grouping of reasoned contributions that ferment in a collective debate[3]. Deliberation, after all, rests not so much on the number and therefore the quantity, but on the consistency of the contributions and, consequently, on the quality. If we want an intense, quality democracy, we will have to work to ensure that the greatest possible number of citizens' contributions are based on reasoned assessments, the fruit of information and personal reflection which, when grouped with others, crystallises in public debate. This is the appropriate scenario and not the mere superimposition of unfounded opinions that does not encourage debate, but rather the impossibility of following the minimum guidelines for collective understanding. The massification of information generated by technology does not help in this regard, so that systems for sifting and cleaning up disinformation will also play a relevant role.

The right use of artificial intelligence can help citizens to assess more clearly the relevance of a decision. AI facilitates a more efficient handling of information, making data available to people so that they can form an informed opinion on the necessity or desirability of certain decisions. The ultimate purpose is to contribute to raising the quality of public debate, in line with the principles of the mentioned deliberative democracy (Tudela 2023).

Consequently, it becomes crucial to promote media literacy and critical thinking so that citizens can discern between truthful and false information and participate in an informed way in public debate, what

[3] Obviously, there are positions in the doctrine that do not encourage the position described above to this extent. Schumpeter, for example, questions whether democracy demands such intense political involvement from citizens.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 155-172
https://doi.org/10.18543/djhr.3193 • http://djhr.revistas.deusto.es/

has come to be called digital literacy. This is no small challenge, of course, but it is the only way for citizens to truly exercise their right to political participation. The capacity for individual analysis is part of each citizen's sphere of freedom, but this analysis can only be sustained if it is based on elements that can be verified. There can be no deliberation, no public debate, when there is false and contradictory information because the reflections derived from it are not conducive to understanding or harmony. On the contrary, disinformation is aimed at confrontation among citizens, the confusion inherent in it only generates a democratic setback because, ultimately, the basic pillars that support the democratic scaffolding are called into question. Without a citizenry aware that not all information that reaches them will necessarily be truthful, and that technology itself is the fuel that *fake news* needs to spread, there will hardly be democratic understanding. An extra effort must be made to observe the veracity of the information that citizens consume and to be prepared for the possibility of having to question certain information that lacks rigour or that overlooks part of reality in order to manipulate or distort citizens' perceptions.

Another aspect to bear in mind with the uncontrolled irruption of technology in the democratic arena is the risk of creating parallel realities in public debate. This occurs when different groups of people have radically different perceptions of reality, based on the information they are exposed to. This can be exacerbated by social networks and recommendation algorithms that filter content to fit user preferences, creating filter bubbles that reinforce existing beliefs and exclude alternative perspectives (Pariser 2011). Its relationship with tribalisation is clear, but beyond grouping people by their common interests, what we are emphasising here is the direct generation of a non-existent reality, artificially created by an uncritical assumption of what is received through new technologies. The nuance is different, and that is why we are expressing it as an element that, although it is clearly related to the above, may be even more worrying due to the perversion of the deception in which citizens can become ensnared. Artificial intelligence algorithms used in digital platforms can contribute to the creation of parallel realities by personalising the content shown to each user, reinforcing their cognitive biases and limiting their exposure to divergent perspectives. Moreover, the spread of disinformation and fake news can be amplified by algorithms that prioritise *virality* over veracity.

To address this problem, it is necessary to encourage diversity of opinion and promote exposure to different points of view, as well as to

develop critical thinking skills to challenge dominant narratives. Strictly speaking, the growth of a society that is individualistic and lacks common reference points is not conducive to channels for debating and listening to different opinions but generating them is certainly the starting point for counteracting these problems. In this sense, deliberation has the potential to transform preferences and interests, which are undoubtedly shaped in a social context, demonstrating that it is not necessary to resign oneself to accepting existing ones as they are or to simply add them up (Cortina 2009).

However, technology can also be used for issues that are beneficial to citizens, such as re-engaging public debate on facts and improving political participation. Thus, to improve the quality of public debate and foster more informed and engaged political participation in the age of social media and artificial intelligence, it is necessary to promote media and digital literacy from an early age, which involves teaching citizens how to assess the credibility of information sources, detect misinformation and engage constructively in evidence-based debates. In addition, effective regulations must be put in place to combat online misinformation and promote transparency in the collection and use of data by technology platforms. Fostering informational pluralism and diversity of voices in the media is also essential to ensure healthy and democratic public debate.

There is no doubt that we are facing two realities that will gradually have to coexist in the social space. Citizens, by the mere fact of being citizens, will continue to participate in public affairs to determine the political reality that organises our coexistence and, for their part, it is impossible to argue that technology will not continue to develop, exponentially moreover, generating new realities from the technological point of view that will have to be integrated into the strictly human reality, with artificial intelligence being the spearhead of this technological development. Thus, observing the unquestionable link between these two elements, reflection should focus on the future that awaits societies when integrating these two issues.

3. **A future perspective on artificial intelligence and its democratic impact**

Artificial intelligence plays a crucial role in shaping public debate and the dissemination of information on social networks. Recommendation algorithms personalise content to suit user preferences, which can create filter bubbles that reinforce existing

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 155-172
https://doi.org/10.18543/djhr.3193 • http://djhr.revistas.deusto.es/

beliefs and exclude alternative perspectives (Pariser 2011). Moreover, AI can also be used to generate and disseminate disinformation more effectively, thereby increasing risks to democratic integrity. Indeed, in the context of elections, AI can be used to coordinate disinformation campaigns (Mateos 2021), mobilise people and use *bots* to spread messages aimed at destabilising a country or creating distrust in electoral processes. The case of Cambodia, where *Meta* recommended suspending the account of former Prime Minister Hun Sen for threatening violence against his political opponents, illustrates how AI can be used to identify and address incitement to violence online.

When reflecting on the future possibilities of a technology as fast-moving as AI, it is wise to be cautious. There are all sorts of approaches (Marichal 2023; López Rubio 2023), some more well-founded than others, but all uncertain. One possibility is that we are at the beginning of a process that will culminate in the complete development of artificial entities through artificial intelligence, which will lead to the reinterpretation of all aspects of human life, including the political aspect. There are, for their part, other approaches that redound to the idea that AI is based on the consumption of data to generate responses and that, when the original data is exhausted, the self-learning mode of these elements will be distorted by having to learn about itself. This would lead to what we see graphically when a photocopy is made of a document which, in turn, has already been photocopied. The logical consequence will be the loss of quality of predictions and behaviour due to the loss of data quality.

Be that as it may, what is important in this respect is to emphasise that the elements integral to the advancement of the new technologies, which are tendentially increasing, will have to lead, in some way, to a substantial modification of the way in which we interpret and manage democratic reality. It will certainly have a direct impact on human decision-making (Papa 2023).

The most pessimistic perspective is that we will be able to be completely guided politically (Rovira 2021), to the point that an artificial intelligence will determine, in a rigorous manner with the study of all available data, what our political choice should be. This is in line with what Innerarity points out when he indicates that "the phenomenon of algorithmic governance is part of a broader trend towards its mechanization" (Innerarity 2020). In doing so, we would renounce our status as citizens. This is the worst approach. But there can also be positive ones based on the same parameters, and AI itself can provide us with the means to find the most relevant information on which to reflect and determine our political adherence. This would

not be a manipulation but, precisely, a valuable tool to dispense with all unverifiable or unverified information. It would shield access to information, or rather disinformation, that would contaminate the political debate. If the elements that pit citizens against each other on the basis of deliberate lies were removed, political debate would be purer and, paradoxically, the very technology that allows for the increased speed of this inappropriate information could be the antidote to prevent access to it.

The latter is the optimistic, positive perspective on how artificial intelligence can curb the advance of disinformation. Generating barriers to entry into the social debate of malicious information will certainly cleanse the public arena of impurities. And another approach, also positive, is the possibility of opening up spaces, especially in terms of time, that allow citizens to interconnect. Artificial intelligence itself can, and indeed must, reduce the workload of citizens. This is a tangible reality that is easily observable in our daily lives. The fear of job losses is notorious, but it is also true that new job opportunities will be provided, as has been the case with previous technological advances. However, many of the activities that have been carried out over the years with a considerable time investment burden will be substantially reduced. Time will be freed up. And that is another scenario conducive to investing time in democratic development, generating spaces for debate and reflection without having to cut hours from other necessary activities. The time freed up by artificial intelligence can be used, without much difficulty, to improve the quality of democratic debate because more time will be available for it.

In close connection with the above, artificial intelligence can digest a quantity of information that is unmanageable for the human intellect, and this is also conducive to democratic improvement. This is based on the principle that we can synthesise information, clarify terms and discriminate information that is not relevant to us with the help of artificial intelligence. Citizen participation can be improved by the contextualisation of information that artificial intelligence can provide. The filtering of this information, not by external or artificial parameters, but with the indications that we each consider, can also be an incentive to establish a more fruitful democratic debate.

At the national level, different legislative reforms must also be addressed to regulate the impact of artificial intelligence systems on democracy. Thus, for example, the Organic Law on the General Electoral Regime (LOREG) should specify the guarantees for the use of personal data through AI tools (Montilla 2023, 42), as well as ensure transparency in the use of these technologies, among other issues.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 155-172
https://doi.org/10.18543/djhr.3193 • http://djhr.revistas.deusto.es/

As we can see, the democratic impact of artificial intelligence will be very significant, but this will depend on the prism from which artificial intelligence is used. If it fits within auxiliary parameters for the political status of citizens, artificial intelligence will improve the quality of democracies. If, on the other hand, it is used as a potential substitute for human reflection and the postulates of artificial intelligence are uncritically followed, we will be treading dangerously on the path to the destruction of democracy as we understand it, in which citizens are the holders of the capacity to decide their own destiny, collectively and subject to political parameters in which peace, freedom and harmony are guaranteed. To opt for citizen subrogation in favour of the postulates of artificial intelligence would, of course, be detrimental to democracy, which is why the importance of technological development as an auxiliary element of human capacities, and not as a substitute for them, should be supported and emphasised.

Conclusions

It is clear that the impact of Artificial Intelligence on democracy will depend on how it is used and how the challenges and risks associated with its implementation are addressed. Thus, the election year of 2024 poses significant challenges for democratic integrity in an increasingly digitised world. The proliferation of disinformation and election-related violence highlights the urgent need to address systemic issues affecting social networks and artificial intelligence. Only through a collaborative and multidisciplinary approach can we ensure that digital platforms fulfil their responsibility to protect and promote democracy in the 21st century.

In any case, as we face this unprecedented election year, it is crucial that social media platforms learn from past mistakes and take concrete steps to protect democratic integrity. This includes implementing more effective measures to combat disinformation, media manipulation and incitement to violence online. In addition, there is a need to promote transparency in the operation of artificial intelligence algorithms and ensure that they are used in an ethical and responsible manner.

Media education and critical thinking are also essential to empower citizens to discern between truthful and false information, and to participate in an informed way in public debate. Government institutions, civil society organisations and the private sector must work

together to develop effective policies and regulations to protect democratic integrity in the age of social media and artificial intelligence.

And, above all other considerations, it is worth emphasising the idea that citizens' participation in public affairs is a human right, a right that perfects the human being by realising it and that allows for the guarantee of other rights of equal importance. The possible alternative to the existence of societies in which citizens participate freely is anything but desirable, so the proliferation of new technologies is to be welcomed, but always bearing in mind that the ultimate guarantee must be one that rests on the protection of human rights, especially one that results in the participation of citizens in public affairs.

In short, it is true that there are many challenges and threats facing humanity in recent years following the emergence and exponential development of AI systems and tools. Legislators and other legal operators have the regulation of these disruptive technologies at the top of their agendas in order not to violate human rights. As is well known, law always lags behind reality, but in the case of artificial intelligence, any attempt to regulate it once it comes into force is doomed from the outset to be obsolete. This does not mean that its regulation should be avoided, but rather that it should be done as far as possible, considering its global nature and with the aim of being a regulation that lasts over time. This has been the premise followed by the recently approved European AI Regulation, which should undoubtedly be accompanied by other national legislative reforms, with the Organic Law on the General Electoral System being one of the first that should be thoroughly reformed to adapt to the digital era, the use of AI and the serious security risk represented by disinformation at election time.

In addition to the threats posed by the impact of AI on democratic processes, positive aspects that reinforce the right to participation have also been highlighted. Thus, it is worth highlighting the capacity of AI-based tools to process large amounts of data, offering clear and detailed analyses that help citizens to make informed decisions; the possibility of creating applications that facilitate public consultations or interactive platforms, increasing citizens' involvement in decision-making or even helping to catalogue and discard false information that clouds the democratic debate.

Without free societies, citizen participation is not possible, so technological development must, imperatively, be developed with this premise in mind.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 155-172
https://doi.org/10.18543/djhr.3193 • http://djhr.revistas.deusto.es/

References

Ackerley, María Isabel. 2017. «La revolución de la democracia: el derrumbe de la democracia representativa y el paso a la democracia participativa, cibernética y a la autogestión». *Eikasía: Revista de Filosofía* 73: 143-162.

Álvarez, Mª Isabel y Federico de Montalvo. 2011. «La teledemocracia como nueva forma de democracia directa en el ejercicio del derecho fundamental de sufragio activo.». In *Derecho y nuevas tecnologías*, coordinated by Ana I. Herrán, Aitziber Emaldi y Marta Enciso, 439-452. Bilbao: Universidad de Deusto.

Balaguer, Francisco. 2023. «Inteligencia artificial y cultura constitucional.» In *Derechos fundamentales y democracia en el constitucionalismo digital*, edited by Francisco Balaguer and Ingo Wolfgang, 41-66. Cizur Menor: Aranzadi.

Carrillo, Nereida, 2022. «Vencer sin convencer: la desinformación en procesos electorales.» *Tribuna Norteamericana* 38: 20-25.

Castellanos, Jorge. 2020. «El derecho humano a participar: estudio del artículo 21 de la Declaración Universal de Derechos Humanos», *Universitas* 30: 33-51.

Castellanos, Jorge. 2024. «Una reflexión acerca de la influencia de la inteligencia artificial en los derechos fundamentales.» In *Ciencia de datos y perspectivas de la inteligencia artificial*, coordinated by Francisca Ramón, 271-300. Valencia: Tirant lo Blanch.

Cortina, Adela, 2009. «La política deliberativa de Jürgen Habermas: virtualidades y límites.» *Revista de Estudios Políticos (nueva época)* 144: 169-193.

Han, Byung-Chul. 2022. *Infocracia. La digitalización y la crisis de la democracia.* Madrid: Taurus.

Innerarity, Daniel. 2020. «El impacto de la inteligencia artificial en la democracia.» *Revista de las Cortes Generales* 109: 87-103.

López Rubio, Daniel. 2023. «Tres reflexiones sobre nuevas tecnologías y democracia directa.» In *Derecho, nuevas tecnologías e Inteligencia Artificial*, coordinated by por Covadonga López Suárez, Ignacio Hernández, Julia Ammerman, Cristina Alonso, Almudena Valiño y Ana Rodríguez Álvarez, 21-26. Madrid: Dykinson.

Marichal, José. 2023. «El peligro de la inteligencia artificial para la democracia», *Anuario internacional CIDOB* 1: 152-154.

Mateos, José Luis. 2021. «La incidencia de los algoritmos en las campañas electorales: un riesgo para la participación política y la democracia.» In *Inteligencia artificial y defensa: nuevos horizontes*, coordinated by Paula María Tomé, Daniel Terrón, José Luis Domínguez, 243-255. Cizur Menor: Aranzadi.

Montero, María Dolores. 2022. «La integración de los Objetivos de Desarrollo Sostenible (ODS) en los nuevos Planes de Acción de Gobierno Abierto.» In *Escenarios de la participación ciudadana: una visión multinivel,* edited by Javier Sierra, Fernando Reviriego y José Tudela, 169-192. Zaragoza: Fundación Manuel Giménez Abad.

Montero, María Dolores. 2023. *Democracia en transición: una agenda para su regeneración.* Madrid: Dykinson.

Montilla, José Antonio, 2023. «Inteligencia artificial y derechos de participación política», *De lege ferenda. Revista de la Facultad de Derecho de la Universidad de Granada* 1: 34-55.

Papa, Anna. 2023. «El uso de la inteligencia artificial en la toma de decisiones públicas: tecnología, política y protección de derechos». In *Derechos fundamentales y democracia en el constitucionalismo digital*, edited by por Francisco Balaguer and Ingo Wolfgang, 209-221. Cizur Menor: Aranzadi.

Pariser, Eli. 2011. *Filter bubble. What the internet is hiding from you*. New York: The Penguin Press.

Ragragio, Jefferson L.D. 2021. «Strongman, patronage and fake news», *Journal of language and politics* 20 (6): 852-872.

Rovira, Joan. 2021. «Algoritmos que predicen y guían nuestras conductas: una reflexión sobre Inteligencia Artificial, libertad y democracia», *Anuario internacional CIDOB* 1: 44-45.

Ruiz Robledo, Agustín. 2018. «El derecho a participar en elecciones libres según la jurisprudencia del Tribunal Europeo de Derechos Humanos.» *Corts: Anuario de derecho parlamentario* 30: 275-305.

Sánchez, Diego. 2006. «Las nuevas tecnologías, el acceso a la información y la participación ciudadana», *Derecho y Tecnología: Revista Arbitrada de Derecho y Nuevas Tecnologías* 8: 209-220.

Soriano, Alba. 2021. «Decisiones automatizadas: problemas y soluciones jurídicas. Más allá de la protección de datos», *Revista de Derecho Público: Teoría y Método* 3: 85-127.

Tudela, José. 2023. «Gobierno, parlamento, democracia e inteligencia artificial", *Teoría y realidad constitucional* 52: 303-333.

Viciano, Roberto and Diego González Cadenas. 2014. «Derecho a participar directa o mediante representantes en el gobierno del propio país: (art. 21.1 DUDH; art. 25.a), 25.b) PIDCP).» In *El sistema universal de los derechos humanos: Estudio sistemático de la declaración universal de los derechos humanos, el pacto internacional de derechos civiles y políticos, el pacto internacional de derechos económicos, sociales y culturales y textos internacionales concordantes*, coordinated by Cristina Monereo, José Luis Monereo and Augusto Aguilar, 335-344. Granada: Comares.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 155-172
https://doi.org/10.18543/djhr.3193 • http://djhr.revistas.deusto.es/

Towards a better protection of human rights through the use of AI and related technologies in budgeting and auditing of public expenditure

Hacia una mejor protección de los derechos humanos mediante el uso de la IA y tecnologías conexas en la presupuestación y control del gasto público

María Amparo Grau Ruiz
Complutense University of Madrid. España
grauruiz@ucm.es
ORCID: https://orcid.org/0000-0002-0482-2816

https://doi.org/10.18543/djhr.3194

Submission date: 08.09.2024
Approval date: 10.12.2024
E-published: December 2024

Citation / Cómo citar: Grau, María Amparo. 2024. «Towards a better protection of human rights through the use of AI and related technologies in budgeting and auditing of public expenditure.» *Deusto Journal of Human Rights*, n. 14: 173-201. https://doi.org/10.18543/djhr.3194

Summary: 1. Human rights and their economic cost: could digitalization (to allow greater financial efficiency) improve their effectiveness? 2. Overview of the current use of AI in the legal-financial field. 2.1. Room for AI progress in the expenditure side. 2.2. Complementarity of technological developments in the private sector. 3. The potential of budgetary digitalization for sustainable development. 3.1. The different technologies and their use along budgetary process and its phases: planning, management, and both internal and external control/scrutiny. 3.2. Trends in the EU. 3.2.1. The adoption of different technologies to improve budgetary processes. 3.2.2. Possible impact of the new regulation on AI. Conclusion. References. Annex.

Abstract: The full potential of many human rights cannot be reached due to the economic costs in their development. The use of artificial intelligence and related technologies in budgetary and audit processes could help in a better allocation of scarce public resources and deliver savings due to better targeting in programming and execution, avoiding irregularities and corruption. When public and corporate organizations automate processes, monitoring should ensure their compliance with regulation or voluntary

commitments affecting environmental, social, and governance criteria. Many funds are granted to support digitalization processes if safeguards related to human rights are respected. The provision of goods and services like health and education is often subject to additional technological requirements. In both cases, an efficient supervision is crucial for fairness, in terms of accessibility and the effective protection of human rights.

Key words: human rights, public expenditure, control, artificial intelligence.

Resumen: El pleno potencial de muchos derechos humanos no puede alcanzarse debido a los costes económicos que conlleva su desarrollo. El uso de inteligencia artificial y tecnologías conexas en los procesos presupuestarios y de auditoría podría ayudar a una mejor asignación de los escasos recursos públicos y suponer ahorros por una mejor orientación en la programación y ejecución, evitando irregularidades y corrupción. Cuando las organizaciones públicas y empresariales automatizan procesos, la supervisión debe garantizar su conformidad con la normativa o los compromisos voluntarios que afectan a criterios medioambientales, sociales y de gobernanza. Muchos fondos se conceden para apoyar procesos de digitalización si se respetan las salvaguardias relacionadas con los derechos humanos. La provisión de bienes y servicios como la sanidad y la educación está sujeta, a menudo, a requisitos tecnológicos adicionales. En ambos casos, un control eficaz es crucial para la equidad, en términos de accesibilidad y protección efectiva de los derechos humanos.

Palabras clave: derechos humanos, gasto público, control, inteligencia artificial.

1. Human rights and their economic cost: could digitalization (to allow greater financial efficiency) improve their effectiveness?[1]

The current international trends link human rights to Sustainable Development Goals in the framework of the United Nations 2030 Agenda, whose implementation was supported with the Addis Ababa Action Agenda approved at the Third International Conference on Financing for Development in 2015. In 2025 Spain will host the Fourth International Conference on Financing for Development and it will address new issues, discussing the reform of the international financial architecture. A recent Resolution adopted by the General Assembly on 22 September 2024 (A/RES/79/1), the Pact for the Future, makes clear the connection among human rights, sustainable development, technological innovation and their financing. It comprises several actions and decisions. Action 7 reaffirms the need to build peaceful just and inclusive societies that provide equal access to justice and that are based on respect for human rights, on rule of law and good governance at all levels and on transparent and effective and accountable institutions. States decide to "promote and protect human rights and the implementation of the 2030 Agenda for Sustainable Development as interrelated and mutually reinforcing". Action 46 stresses: "We will ensure the effective enjoyment by all of all human rights and respond to new and emerging challenges", recalling that "the Sustainable Development Goals seek to realize the human rights of all". Action 30 is entitled: "We will ensure that science, technology and innovation contribute to the full enjoyment of human rights by all". In paragraph 50 is added: "We will deepen our partnerships with relevant stakeholders, especially the international financial institutions, the private sector, the technical and academic communities and civil society, and we will ensure that science, technology and innovation is a catalyst for a more inclusive, equitable, sustainable and prosperous world for all, in which all human rights are fully respected". A Global Digital Compact has been included as Annex I in the Pact for the Future. First, it mentions that "digital technologies are dramatically transforming our world. They offer immense potential benefits for the well-being and advancement of people and societies and for our planet. They hold out the promise of accelerating the achievement of

[1] This work has been carried out as PI in the framework of the research project: "Developing SustAI'nAbility" (FEI-EU-23-02). The author thanks Cassandra Bouzi for her help in searching some useful materials.

the Sustainable Development Goals". The States recognize "the need to identify and mitigate risks and to ensure human oversight of technology in ways that advance sustainable development and the full enjoyment of human rights". Objective number 3 is to "foster an inclusive, open, safe and secure digital space that respects, protects and promote human rights". Regarding principles, the Compact "is anchored in international law, including international human rights law. All human rights, including civil, political, economic, social and cultural rights, and fundamental freedoms, must be respected, protected and promoted online and offline. Our cooperation will harness digital technologies to advance all human rights".

Despite the aim of promoting, protecting and fulfilling all human rights, the full potential of many basic human rights often cannot be reached because of the economic costs in their development (Grau 2020, 175). The classic design and implementation of budgetary policies should be improved for a fairer allocation of resources, and digitalization could offer a chance to introduce sound changes in both areas. Exponential progress in technological developments like big data analytics, artificial intelligence (AI), digital platforms, robotic process automation, distributed ledger technologies and satellite imagery could lead to greater efficiency and respect of fundamental rights in policy, law making and implementation of law. There are consequences of actions or omissions resulting from the poor quality of policies and legislation (e.g. billions of euros are lost annually due to missed or delayed policies and legal reforms). Policy, law making, and regulation can lead to regulatory failure causing inefficient allocation of resources or unintended redistribution, among other effects (Maciejewski 2024, 38, 53-54). Thus, an ex-post quantified evaluation of legislation (and budgetary choices) needs to be applied consistently. An 'intelligent' performance-based policy and law making could be implemented, in addition to compliance control to avoid losses of budgetary resources (Grau 2023, 72-91).

In the European Union (EU), the European Parliament acts as a co-legislator with the Council and adopts laws for over 350 million European citizens, but it also acts as budgetary authority. It has been frequently caring about a sound financial management of EU funds. Silos approach to goals and lack of coordination of instruments, lack of addressing efficacy of regulation and biases of actors with recurring concerns as to their objectivity and the role of pressure groups, have been present in the discussion on EU policy, law making and regulation since decades. Nowadays there are isolated ICT tools aimed at addressing these concerns (Maciejewski 2024, 37, 41).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

Worldwide many governments have opportunities to deliver substantial productivity gains and transform public services to deliver better outcomes for the taxpayers, but public and civil servants should have the tools, information and skills they need to use AI, so that the public trusts the government's responsible use of AI (National Audit Office 2024, 1). Recently, algorithmic approaches have emerged. They can focus on how specific expenditures (budget inputs) are processed to generate economic, political, and social outcomes (outputs). This could serve in the planning stage of the expenditure allocation process and the distribution of public spending to increase GDP, decrease inflation and reduce inequalities. By offering criteria to leverage multiple or conflicting objectives, this type of approach could complement or substitute other analytical techniques used to make decisions about budget allocations. It could even bring some degree of rationality to the budgetary process with evidence to support best practices and understanding of the data used for specific government programs (through simulations). An analysis of all available data (regardless of its distribution, size, or format) makes it easier to detect which expenditure allocation strategies have been (or not) successful in the past and helps dynamic allocation. Therefore, public servants —required to use and account for every dollar— should be always on the lookout for new capabilities and tools to help them make better decisions. AI can make more decisions more cheaply and faster, nevertheless one cannot overlook other aspects like the necessary computational capacity, and the lack of algorithmic transparency —that might result in bias, omissions, and errors (a quite sensitive issue where there are many self-interested actors involved) (Valle-Cruz et al. 2022, 13)[2].

Obviously, any technological innovation and development at the expense of human rights is counterproductive. A right-based approach is essential to assess AI progress. It is said that the right to development will be breached if there is lack of effective and meaningful participation through which individuals and peoples contribute to, and enjoy economic, social, cultural, and political development, in which all human rights and fundamental freedoms can be fully realized (Mahmutaj

[2] They explain that "despite the black box inherent in AI algorithms, such techniques can bring some degree of rationality to the budget process, which is not only technical, but also political in nature", recognizing that "an automated AI system that makes technical decisions to be accepted by politicians or other decision-makers is still far from being technically possible and politically feasible". Bias may exist in the datasets too and some policies may be effective for certain nations and contexts but may lead to failure or harmful biases for others.

forthcoming). However, government officials can use different tools and resources (like the OPSI Toolkit Navigator) to help identify and engage with users and individuals who may be affected by an AI system in order to better understand their point of view (Berryhill et al. 2019, 108).

2. **Overview of the current use of AI in the legal-financial field**

2.1. *Room for AI progress in the expenditure side*

When adopting new technologies, governments need to determine the appropriate trade-off between strong controls and experimentation and risk, based on the relative costs and benefits, as it happens in the revenue side of the public financial activity (Grau 2022a, 325; Martín López 2023, 44). Accordingly, the public sector leaders should assess the nature of interactions in the AI systems for which they are responsible and determine whether they are appropriate. It is worth noting that algorithms need to be trained to provide a viable service, and there is always a chance that an AI will not perform as intended —even with an unbiased algorithm— and controls will not reduce risk to zero. Nonetheless, postponing AI deployment will delay the realization of the benefits it can bring. Similarly, existing decision-making processes are unlikely to be completely accurate and unbiased (Berryhill et al. 2019, 109).

Eventually, in the digital age, it is not a question if AI will proliferate, but when, how, and by whom. It can impact agencies' immediate annual spending, but also the government's long-term capacity to serve the people. For example, in the US federal agencies are harnessing AI, workflow automation, and other advanced technologies for budget forecasting and planning. Workflow automation can integrate diverse data sources by linking and pulling information from different systems, updating them all in real time. AI capabilities can improve this process with gap analysis and automated recommendations in line with the agency's mission. Applied at scale, they can reduce decision time and increase transparency, discovery, and traceability of data across programs. These agencies aim at improving strategic alignment, accelerating mission success, and preserving — and potentially expanding — their budgets (Fedeli 2024, 1)[3].

[3] "Federal managers must pore through mountains of data across thousands of programs. They must make sound decisions despite persistent challenges: potentially duplicative projects, unseen dependencies, time-consuming data calls, stove piped data

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

2.2. *Complementarity of technological developments in the private sector*

Government agencies can exploit external sources of information to better achieve their missions. In fact, many industries produce amounts of data, regularly in machine-readable formats (Berryhill et al. 2019, 117). The collection and generation of data by the private sector provide opportunities for the use of AI. This phenomenon is reinforced when financial and sustainability reporting is demanded (Grau 2022b, 61).

Some tools developed in the context of companies, can be valuable for the authorities in charge of the public budget. Thus, enterprise decision management platforms with AI-augmented workflow automation help visualize data in new ways, revealing how money flows through complex organizations and leads to on-the-ground results (Fedeli 2024, 1)[4]. This may be relevant for budget planning. In the same vein, the use of digital twins —virtual representations of people, places and things— is emerging. They can take advantage of major advancements in machine learning, analytics, AI and augmented reality/virtual reality (Preut et al. 2022, 1).

As it happens with the real costs of most regulatory programs —that are borne, not by the regulators but by the firms and individuals who have to comply with the regulation (Maciejewski 2024, 37)—, the budgetary authorities many times can freely enjoy certain economic savings due to previous technological advancements in the private sector.

---

sources, legacy tools and formats, and low visibility into outcomes" [...] Program managers can streamline data calls about cost, schedule, and performance. All three can link funding, project, portfolio, and capability data for visualizing alignment mapping of budget flows, from resource sponsors down to the minutiae of project execution". [...] "Empowering managers to make better decisions faster not only improves federal agencies' immediate annual spending, but also the government's long-term capacity to serve the American people. Many federal leaders seek predictive tools for divestment decisions —figuring out what stays and what goes to achieve optimal return on taxpayer money. As a few simple clicks reveal duplicative or deprioritized efforts, users can chart new pathways to free up resources, illuminate modernization pathways, and scale up innovation".

[4] The following features are useful: similarity analysis and contextual search (because it can reveal hidden relationships across thousands of programs); command view (single interface that integrates all systems and data); machine learning (through AI suggestion and user interaction, while keeping the historical "why" behind every decision); fast setup (new capabilities need to work with existing systems without rip and replace).

3. The potential of budgetary digitalization for sustainable development

Successful risk-informed Sustainable Development Goals (SDG)-financing solutions depend on prudent public financial planning and management, and budget execution. Finite public funds explicitly contend with diverse SDG priorities, and sometimes funds spent to advance an SDG implicitly negatively impacts outcomes on others. Policymakers should determine how these risks interact and how the expenditures affect SDG performance. They can use 'Budgeting for the SDGs' (B4SDGs) for budget planning, delivery, execution, and evaluation cycles. This tool supports the development of evidence-based medium-term revenue and expenditure strategies with a view to optimize public spending efficiency, reduce opportunity costs and wastefulness, and enhance the effective use of domestic public resources (United Nations 2022, 1).

The iBiT is an artificial budget intelligence powered toolkit. It has been developed by ESCWA to amplify the returns on public spending and optimize public expenditure efficiency. This toolkit aims at maximizing SDG performance across national targets, it overcomes fiscal space (Navarra et al. 2024, 30)[5] limitations, and captures SDG-budget synergies and trade-offs in different country contexts (United Nations 2023, 1). The iBiT provides insights on the cumulative contribution of public spending on SDG performance, the budget lines advancing or regressing the SDGs, how optimal a change in allocation is, how impactful the budget spending is, how to optimize it and meet constitutional thresholds, how much more could the SDGs progress if additional financing was optimized, etc.

However, to ensure the efficiency of programs and policies it is necessary to integrate a systematic and consistent data collection phase in order to monitor and evaluate the commitment and the impact of a certain issue, for example, regarding gender equality. 'Gender-sensitive budgeting' or 'gender budgeting' means gender mainstreaming of the entire budgetary process with a view to incorporating a gender equality perspective to all decisions on revenue and expenditure. This has a fundamental impact on inclusive

[5] "While there is no commonly agreed definition, fiscal space relates to 'the financing of policies conducive to the development of a country […] both in its narrow sense, as a redefinition of the fiscal rules to which sensible fiscal policy has always been subject, or in broader term as a full-blown set of policy actions for development' (Aguzzoni, 2011; Roy et al., 2009)".

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

and economic growth, fostering employment, reducing poverty, addressing ageing population and increasing GDP (European Parliament 2024, 29-30)[6].

3.1. *The different technologies and their use along budgetary process and its phases: planning, management, and both internal and external control/scrutiny*

In every jurisdiction, each level of government has its own approach to the budgeting process, but most basically follow four steps: prepare, approve, implement, and audit (Johnston 2023c, 1). Transparency in public sector budgeting helps deter corruption and fraudulent use of public funds everywhere. Currently, priority-based budgeting (PBB) practices allow visibility into the budgeting process, which is very useful for accountability, inclusiveness, trust and quality purposes. As the world becomes increasingly technology focused, budgeting processes must adapt to keep up, but without diminishing their transparency.

Cloud budgeting software makes it easier to track metrics that show how well program funding aligns with priorities or adjust it as needed. Asynchronous collaboration capabilities allow team members and decision makers to work simultaneously on a budget and collaborate across departments. This shortens the budget timeline, streamlines processes, and increases productivity (Johnston 2023a, 1).

The detailed analysis of AI techniques as a tool to support government decision-making in the specific function of public budgeting is relatively scarce, despite being one of the most important functions of government. They can provide ideas to classify public

[6] A 2023 briefing on 'Gender budgeting in the Member States', produced by the Policy Department for Budgetary Affairs for the FEMM, BUDG and CONT Committee found that 12 countries have introduced gender budgeting; nine countries do neither practise gender budgeting nor consider introducing it (the reason given is that often the country's gender equality policy is seen as sufficient); three countries have not yet introduced gender budgeting but are discussing its potential usefulness. The Special Report of the European Court of Auditors 'Gender mainstreaming in the EU budget: time to turn words into action' concluded that the EU's budget cycle did not adequately take gender equality into account; the Commission made limited use of sex-disaggregated data and indicators, and published little information on the EU budget's overall impact on gender equality. The Commission has been developing a methodology to track all EU spending programmes' contributions to gender equality. The methodology, a work in progress, was implemented in a pilot in the 2023 Draft Budget (DB2023) and again for the 2024 Draft Budget (DB2024) and it is currently being implemented in reporting exercise for the 2025 Draft Budget (DB 2025).

budgeting allocations to different programs and policies and identify some opportunities and scenarios that ultimately government leaders can assess. For example, AI modeling technology can help move quickly through customized PBB implementation when governments lack the time, internal staff, and resources. The existing departmental budget data —including personnel, non-personnel, and operational line items— are used to create an inventory of the programs, then AI modeling allows to predict how much of the strategic budget is applied to a specific program also helps decision-makers score programs. After analyzing huge datasets and pinpointing where budgetary dollars can be cut or reallocated, AI modeling provides insights by communicating the budget programmatically (rather than departmentally by line item) and adds visibility into what choices are being made to provide services to residents and why. It can also identify ways to create new revenue streams to fund high-priority programs, and show opportunities to outsource programs to private companies, and potential partnerships with other government entities to share resources. Evidently, better understanding of the actual costs of a service helps develop and manage budgets more efficiently and increase their impact on community priorities (Resource X 2023, 1). For example, in North America, The Government Finance Offices Association's (GFOA) Rethinking Budgeting initiative helps state and local government leaders better meet community needs by introducing improvements, like PPB, new technologies and budgeting software, and best practices that support successful community outcomes (Johnston 2023b, 1)[7]. Government budgeting software helps governments switch from limited-visibility line-item decisions to data-backed outcomes that ensure efficient and effective resource allocation aligned with the priorities of their communities. This software offers a level of granularity that allows to see the full impact of every budgeting decision across all departments. With this holistic view decision makers can forecast needs and measure costs versus value more effectively (Johnston 2022b, 1)[8].

[7] GFOA was founded in 1906 to facilitate positive change and advance excellence in public finance. The association, which currently comprises more than 20,000 members, represents federal, state/provincial, and local finance officials across the United States and Canada.

[8] Traditional budgeting takes more of a rigid, all-or-nothing tack and is based on historical budgets, in which past decisions are frozen past the point they are affordable or relevant. The priority-based approach describes the budget in terms of programs. This is more relevant to how residents and elected officials experience government services. The focus of priority-based budgeting is on accountability for the results that formed the basis of a program's budget allocation, not whether the program stayed

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

Any allocation of funds has a direct impact on the level of satisfaction of human rights, so technological improvements in the tools to trace and control their use can positively revert to people, just by assuring that they reach their intended beneficiaries. Besides, it may have an indirect impact if these tools produce an increase in the available funding due to greater efficiencies. A discussion could be opened about the convenience to reinvest any savings made from the use of one technology into other forms of IT investment, as well as reskilling programs, instead of allocating that additional revenue to further develop a specific right. Somehow, a sort of multiplying effect is expected. Here, one could even explore simulating different scenarios with AI models to understand potential relationships between public budget expenses and other social and economic outcomes useful for government decision-making (Valle-Cruz et al. 2022, 12)[9].

The adoption of sophisticated IT tools is also regarded as a possible solution that could improve the efficiency of the assurance process and also the audit quality. Digitization allows easier and quicker access to important documents of an operation or intervention during verifications and thus, reduction of the number of controls. Therefore, the audit bodies could more easily use the results of the first level controls (if existing) and retrieve from the IT system any relevant document that is needed to consolidate the audit results. Patterns or structural issues would be much easier to be detected (Malan and Dimauro 2022, 71, Grau 2023, 72).

3.2. *Trends in the EU*

Budgetary authorities have increasingly used new digital technologies to protect the EU budget, because the misuse of EU funds

within spending limits regardless of the outcome. Rather than across-the-board cuts, this approach reduces funding based on the value of the program or service.

[9] "The findings suggest enhancing the allocation of public spending, improving public debt and public expenditure, fostering the investment in agriculture, education, and public health, and implementing strategies to address the problem of unemployment to boost economic growth, decrease income inequality and reduce inflation". These authors use the multilayer perceptron and a multi objective genetic algorithm to analyze World Bank Open Data from 1960 to 2019, including 217 countries. They also propose a hybrid AI approach based on the learning capacity of artificial neural networks. They recognize some limitations in their approach because it does not consider aspects inherent to the budgeting process, such as political, economic, and even corruption-related factors.

remains a serious problem (European Parliament 2021, 22)[10]. However, there has not been a broad and consistent deployment of data-driven technologies in budgetary control across the EU due to differences in national control strategies and systems, regulatory frameworks, investment capacity, digital competences and political priorities between Member States. In the end, consistent adoption of data-driven technologies might support the harmonization of control practices and standardization of reporting methods.

Many of the national recovery and resilience plans have included reforms and investments aimed at introducing or improving e-government services (Collovà et al. 2024, 2)[11]. An initial analysis (in six Member States) of measures relating to digital public services in the national recovery and resilience plans shows that the top two categories in terms of budget allocated are the general 'IT solutions, e-services and interoperability' categories, including at regional level (ranging from 33 % in Italy and France to 89 % in Spain), followed by health (ranging from 6 % in Spain to 66 % in France) (Lilyanova 2024, 9-11)[12]. However, this general category encompasses several policy areas, including health and justice. Conversely, interoperability[13] and

[10] Member States reporting a total of 12,455 irregularities, amounting to EUR 1.77 billion, in 2022. For example, Arachne is a risk-scoring tool used by managing authorities on a voluntary basis to detect risks of fraud and irregularities in the use of European Structural and Investment Funds. However, Arachne is limited by low awareness of the tool, privacy concerns, a high administrative burden, limited accessibility, inaccurate risk scores, and a high number of false positives. The Early Detection and Exclusion System is a database allowing EU bodies to flag financial risks posed by (potential) recipients of EU funds. It does not apply to funds under shared management, but a targeted extension to all management modes from 2028 is expected. The Irregularity Management System is a database within which Member States report irregularities in the management of EU funds, however, its utility is limited by the substantial variation in reporting practices across Member States. The future of digitalisation in budgetary control Executive Summary, Study for the CONT Committee, 1-6. Data-mining tools can indeed make monitoring system more efficient and able to detect fraud and mismanagement of public funds. Arachne processes and analyses data of two million beneficiaries and crosses it with information from external databases that contains information on more than 210 million companies and 120 million people that are behind those companies. Further information on Arachne is available at: https://op.europa.eu/en/publication-detail/-/publication/71c53825-fbb9-11e5-b713-01aa75ed71a1/language-en.

[11] The Commission has created the Innovative Public Services Observatory, which analyses trends, identifies good and bad practices and assesses the impact of new technologies, such as AI, on the public sector.

[12] EPRS initial analysis of measures relating to digital public services in the national recovery and resilience plans of Italy, Germany, Spain, France, Greece and Slovenia.

[13] The level of interoperability of network and information systems supporting digital public services in the EU is still insufficient. This leads to limited digital public

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

the regional dimension are often a key feature of other policy areas, including education and health. The third category is transport. Justice and digital identity are allocated less funds, and are, for instance, absent in Germany and Spain.

3.2.1. The adoption of different technologies to improve budgetary processes

Nowadays, big data analytics, AI, machine learning (ML), natural language processing, deep learning, large language models (LLM), robotic process automation (RPA), blockchain, and satellite imagery are being used by EU Member States to improve budgetary control practices (Rampton et al. 2024, 29-48, 56-64). They are mainly applied to information management of large volumes of data and risk-scoring. Of course, developing AI-powered tools is costly and takes time, requiring constant updates (e.g. they may not be able to capture new indicators of fraud that have not been defined based on auditors' experience, and may generate false positives).

Big data analytics and data mining can facilitate access to data, risk-scoring, interoperability between institutions and harmonized data collection, verification and analysis. Generative AI/LLMs can allow for the summarizing of large datasets, automatically correct, standardize and organize data, allow cross-referencing against other sources, and generate written reports. Platforms using LLMs allow to process large bodies of complex data and text and to retrieve relevant information instantly. However, there is a risk of high levels of inaccuracy in the output of LLMs, high energy consumption and limited scalability.

NLP applications can help those managing and auditing funds 'chat with their docs' with custom-built chatbots. Internal chatbots are currently piloted in a few audit institutions in the US and in Europe (e.g. Cequence for public procurement officials in Czechia and Slovakia). Public-facing chatbots are not yet used in the field of

---

services and causes a number of problems for citizens, organisations and businesses, as well as for public authorities themselves. The interoperable Europe act introduces an obligation to share certain interoperability solutions (such as open-source software) and data between public sector bodies, institutions, bodies and agencies of the Union, with a focus on removing unnecessary burdens (such as legal, organisational, semantic and technical obstacles). The aim is to save citizens, businesses, and the public sector itself, money and time. Public sector bodies and institutions, as well as EU agencies or bodies, would have to evaluate the impact of changes in information technology systems on EU cross-border interoperability.

budgetary control, they could help those managing and auditing public funds communicate with citizens in the future (LLM-powered chatbots could bridge information gaps about EU funds, like Bürokratt, Estonia's 'one-stop-shop' chatbot to ease the burden of applying for small organizations with lower budgets. It could also save managing and paying authorities time as clear instructions will help beneficiaries submit all relevant pieces of information in a timely manner). NLP is used in combination with machine learning techniques to detect signs of irregularities, or patterns that indicate risks of fraud in audit files.

RPA automates repetitive or time-consuming, rule-based tasks that require a high degree of accuracy. This allows the audit teams to focus on higher-value or more complex tasks. It can help to make rapid and effective improvements and to meet strict deadlines and respond quickly. Its constant operation enhances productivity. It can enable web-scraping for data extraction, verification and reporting.

In the context of budgetary control, RPA technology is used to automate data extraction from various sources and consolidation into a central system, reconciliation processes by matching data to ensure accuracy, report generation (budget summaries through formatting data into predefined report templates), audit-trail creation by tracking and recording changes made to financial documents, compliance checks, and budget monitoring to issue an alert when there are deviations from the planned budget.

The use of RPA in budgetary control is limited by its inability to automate complex tasks that require advanced decision-making as well as the necessary shift in organizational culture. RPA cannot learn from past experiences or adapt to new situations without human intervention. AI can help RPA automate tasks more fully, handle more complex data, and find patterns or extract meaning from images, text or speech. In turn, RPA can enable AI insights to be actioned faster without having to wait for manual implementations. The newly combined concepts of Intelligent Automation (IA) and hyperautomation are capable of streamlining numerous procedures, including procurement and payment processes[14]. They can contribute

[14] IA describes the combination of RPA, AI and other related automation technologies. IA technology can analyse data, learn from patterns, make decisions based on historical data, and perform tasks that traditionally required some level of human judgment or intervention. One example of the application of IA technologies in practice is the Intelligent Document Processing, which uses IA to extract, process and validate data from images and other files where data often appears in an unstructured format. [...] The term 'hyperautomation' describes the evolution or extension of IA across a wider range of organizational processes with the aim of creating an

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

to a more effective, efficient management and control of EU funds (especially those under the shared management mode), better resource utilization, reduced administrative burden and enhanced service delivery. Gravitation towards low-code/no-code tools such as RPA is likely in audit institutions, as these solutions empower non-technical users to implement process improvements swiftly.

ML can enhance risk-scoring, strengthen prevention and detection of irregularities, identify weaknesses in control systems and increase understanding of factors causing anomalies. In the last ten years, both NGOs, CSOs, and government agencies have started using machine learning technologies to build 'red flagging' tools. Most of them use manually defined indicators. Researchers, auditors or NGOs examine past cases and identify patterns of fraud in subsidies and/or public procurement contracts. New approaches use unsupervised machine learning algorithms to learn which patterns are associated with higher risks of fraud and corruption. ML algorithms could include an internal chatbot allowing auditors to ask questions about any audit files and be pointed to the relevant file, enabling them to quickly fact-check its answer. Auditors in Belgium, Norway, Portugal, Spain and Sweden are developing tools that will use ML technology to find indications of fraud in large documents of audit data and explore ways to potentially move away from a sample-based auditing process to a 100% AI check. In Massachusetts and New York AI-powered risk-scoring tools are already in use. EU and national-level risk-scoring tools using ML will be key components of the fraud prevention and detection strategy in the future. They have proven their potential to improve fraud detection rates, to recover costs, and to protect national and EU budgets. Challenges in developing national-level risk-scoring tools will depend on the situation in the Member States especially regarding data availability and data interoperability. One source of information on irregularities is the EU-wide dataset of irregularities stored in the Irregularity and Management System (IMS). Any red flagging tool using machine learning in any Member State could be then trained using the IMS data. The more this tool is used, and the more information it contains, the more valuable it will be as a source of data for EU-wide risk-scoring tools.

Digital platforms can facilitate information and knowledge sharing, development of joint initiatives with verification of results between

---

interconnected and automated workflow across the organisation. Hyperautomation of complex business processes that involve both structured and unstructured data has emerged as a significant trend in recent years.

authorities, and harmonized approaches to auditing and control. This can enhance the efficiency, speed, accuracy, and quality of budgetary control, as well as fraud detection activities.

In the context of budgetary, digital platforms are instrumental in enhancing the efficiency, transparency, and effectiveness of budgetary control practices. These tools act as centralized budgetary information and data repositories, making the data readily available to relevant stakeholders Control audit team members can feel more motivated as they actively participate in decision-making and problem solving. All parties involved in budgetary control have access to the most up-to-date financial data and work with the same data, reducing discrepancies. Transparency and access to real-time data is crucial for accurate budget tracking and forecasting. By incorporating new AI based technologies and/or Robotic Process Automation, the platforms can enable automation of manual tasks. However, they also bring issues related to data security, privacy, and interoperability, especially in a multilingual and multi-jurisdictional context like the EU. Strategic implementation and continuous development of both technology and human resources as well as the standardization of data formats and language play are fundamental in order to reap the benefits of digital platforms in budgetary control.

Blockchain can enable the traceability and identification of transactions, streamline data collection and storage, and support efforts to combat fraud. All blockchain transactions are permanently recorded, visible to everyone in the network, and almost impossible to tamper with. By eliminating intermediaries, blockchain technology offers the potential to eliminate opportunities for corruption. It is not yet widely used in budgetary control, there are pilot projects to curb corruption in public procurement, e.g. in Brazil, Columbia, Nigeria, Peru, Rwanda and South Africa. The EU already introduced the European Blockchain Services Infrastructure. In the future, the network could be used to track and record payments in any EU fund and reduce opportunities for intermediaries to divert payments or for beneficiaries to use payments in in ways that are not intended. The EU could also develop a private and permissioned grant management and/or public procurement system based on blockchain technology. But there are challenges like high set-up costs, data protection concerns and high energy consumption.

Satellite imagery is being used for budgetary control, mainly in the Common Agricultural Policy. A new 'checks by monitoring' approach combines satellite data with the information provided by farmers. The EU's Copernicus Sentinel satellites provide frequent and

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

high-resolution images and data to paying agencies. The paying agencies use big data analytics and machine learning algorithms to assess the type of crop and the activities on each declared parcel for each aid scheme. Then, they visualize compliance on digital maps of the respective fields, divided into small parcels. Any parcels the machine learning algorithm assesses as compliant are colored in green. Any parcels it assesses as non-compliant are colored in red. Parcels that require further processing (for instance, because there are indications of potential non-compliance or because results are inconclusive) are colored in yellow. The new checks are automated and continuous: paying agencies monitor agricultural activity throughout the year and check them against the information they receive from the farmers. The new system allows paying agencies to monitor all agricultural parcels in the respective region, they only carry out field visits if the satellite-based monitoring process is inconclusive and if the financial impact of non-compliance exceeds a certain threshold. Paying agencies have more leeway to warn farmers in the case of non-compliance. Farmers also receive useful data to increase the productivity of their farm. However, the overall take-up by paying agencies is still low. Imaging and AI could transform multiple EU monitoring and budgetary control systems in the future.

The main benefits of all the above-mentioned technologies are summarized in table 1.

Table 1.

Benefits of new technologies in budgetary control

Technologies	Benefits of new technologies in budgetary control
Big data analytics and data mining	— Easier and quicker access to important data during verifications. — Enhanced risk-scoring and thus detection of irregularities / fraud. — Cross-border organisation/institutional interoperability. — Harmonisation of data collection, verification and analysis. — Streamlining the audit process and improvement of the audit trail.
Machine learning	— Enhanced risk-scoring, accuracy of red flags and identification of patterns. — Stronger prevention and detection of irregularities/fraud/ corruption in the EU expenditure. — Identification of weaknesses in the national control systems for EU funded programmes. — Better understanding of the explanatory factors leading to a situation/anomalies.

Technologies	Benefits of new technologies in budgetary control
Generative AI/LLMs	— Possibility to summarise large amount of data and information. — LLMs can be used to automatically correct spelling errors, standardise formats, and organise data into structured formats like tables or spreadsheets. — LLMs can be used to cross-reference data against other sources to verify accuracy and reliability. — LLMs excel in generating written content - can automate the creation of reports, summaries, and documentation by structuring collected data into coherent narratives, following specified templates or guidelines.
Robotic process automation	— Web-scraping tools or external APIs can be used for data extraction, verification and reporting thereby streamlining the entire control and assurance process. — Automate repetitive and time-consuming tasks to allow authorities to focus on strategic tasks.
Digital platforms	— Sharing of knowledge by Member States in the use of effective IT tools. — More effective sharing of management verification results. — Reduce gold-plating due to the introduction of unnecessary national / regional rules.
Blockchain	— Traceability and identification of operations and transactions. — Capacity to streamline data collection and to store immutable and reliable data. — Facilitate tax administrations' efforts to deter and combat tax fraud (including cross-border).
Satellite imaginery	— Deep learning image classification algorithms on high-definition satellite imagery to monitor the quantity as well as the quality of crop yield and to check applications for EU funds. — Can be leveraged for budgetary control purposes to verify the quantity and quality of agricultural output funded by the CAP funds and detect anomalies.

Source: Rampton et al. (2024)[15].

Rampton et al. (2024) have stated: "leveraging the strengths of both new and existing technologies can lead to synergies that address a wider range of challenges and requirements within budget management and control processes. The aim is to connect disparate tools and platforms to create a unified ecosystem that supports the

[15] From the same source, there are additional tables 2, 3 and 4 in the annex synthetizing respectively benefits and limitations of risk-scoring tools, digital platforms and RPA.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

entire lifecycle of budget management, from planning and allocation to execution and reporting". Their report contains recommendations: continue to enhance existing EU tools for budgetary control[16]; promote awareness of and training in their use; discuss their compulsory use; consider pilot projects developed on a transnational basis to explore the possibilities for applying new data-driven technologies to budgetary control; support mutual learning, good practice might inspire budgetary authorities to adopt new tools; consider defining common standards for the use of new technologies in budgetary control accompanied by a code of conduct for their proper and fair deployment; assess the costs and benefits before deploying new technologies; carry out regular horizon scanning to identify potential new technological developments suited for application to budgetary control and share information about such developments with budgetary authorities at EU level and in the Member States.

3.2.2. Possible impact of the new Regulation on AI

The Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on AI[17] could relatively impact on some of the described tools for improving budgetary processes. Their benefits and challenges should be assessed with the following criteria —extracted from the text of this legal instrument— in mind.

In principle, as stated in its recital (4), "by improving prediction, optimizing operations and resource allocation, and personalizing digital solutions available for individuals and organizations, the use of AI can provide key competitive advantages to undertakings and support socially and environmentally beneficial outcomes". However, recital

[16] This includes expanding Arachne to all management modes, integrating advanced technologies, ensuring interoperability with other tools, addressing privacy concerns, and enabling faster checking of operators against more up-to-date and comprehensive data cases. The IMS could be improved by introducing consistent thresholds for reporting cases of fraud and providing more up-to-date information. For Arachne, training would include how to use all the different functionalities; for the IMS, thresholds for reporting cases of 'suspected' and 'established' fraud.

[17] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on AI 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 (AI Act) (Text with EEA relevance), OJ L, 2024/1689, 12.7.2024, https://eur-lex.europa.eu/eli/reg/2024/1689/oj

(58) stresses that the use of AI systems deserves special consideration when it affects "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 [...] 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". Still the public administration can benefit from a wider use of compliant and safe AI systems, if they do not entail a high risk to legal and natural persons. Anyway, sandboxes are envisaged in the Regulation and they might be an option to experience new capabilities with limited pilot projects[18].

The recital (20) literally reads: "In order to obtain the greatest benefits from AI systems while protecting fundamental rights, health and safety and to enable democratic control, AI literacy should equip providers, deployers and affected persons with the necessary notions to make informed decisions regarding AI systems. Those notions may vary with regard to the relevant context and can include understanding the correct application of technical elements during the AI system's development phase, the measures to be applied during its use, the suitable ways in which to interpret the AI system's output, and, in the case of affected persons, the knowledge necessary to understand how decisions taken with the assistance of AI will have an impact on them".

Finally, recital (96) adds: "In order to efficiently ensure that fundamental rights are protected, deployers of high-risk AI systems

[18] Article 59.1. In the AI regulatory sandbox, personal data lawfully collected for other purposes may be processed solely for the purpose of developing, training and testing certain AI systems in the sandbox when several conditions are met. They include "(a) AI systems shall be developed for safeguarding substantial public interest by a public authority or another natural or legal person and in one or more of the following areas: [...] (v) efficiency and quality of public administration and public services".

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

that are bodies governed by public law [...], should carry out a fundamental rights impact assessment prior to putting it into use. [...] The aim of the fundamental rights impact assessment is for the deployer to identify the specific risks to the rights of individuals or groups of individuals likely to be affected, identify measures to be taken in the case of a materialisation of those risks".

The impact assessment should be performed prior to deploying the high-risk AI system and should be updated when the deployer considers that any of the relevant factors have changed. The impact assessment should identify the deployer's relevant processes in which the high-risk AI system will be used in line with its intended purpose, and should include a description of the period of time and frequency in which the system is intended to be used as well as of specific categories of natural persons and groups who are likely to be affected in the specific context of use. The assessment should also include the identification of specific risks of harm likely to have an impact on the fundamental rights of those persons or groups. [...] deployers should determine measures to be taken in the case of a materialisation of those risks, including "for example governance arrangements in that specific context of use, such as arrangements for human oversight" according to the instructions of use or, complaint handling and redress procedures, as they could be instrumental in mitigating risks to fundamental rights in concrete use-cases. [...] Where appropriate, "to collect relevant information necessary to perform the impact assessment, deployers of high-risk AI system, in particular when AI systems are used in the public sector, could involve relevant stakeholders, including the representatives of groups of persons likely to be affected by the AI system, independent experts, and civil society organisations..."

The detailed provision for Fundamental rights impact assessment for high-risk AI systems is Article 27.

In ANNEX III, pursuant to Article 6 (2), this area is listed considering high-risk AI systems: "5. Access to and enjoyment of essential private services and essential public services and benefits: (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 is important to note that, upon fulfilment of certain conditions, Article 7.1. allows the Commission to adopt delegated acts in accordance with Article 97 to amend Annex III by adding or modifying use-cases of high-risk AI systems.

Article 77.1 clarifies that "National public authorities or bodies which supervise or enforce the respect of obligations under Union law protecting fundamental rights", including the right to non-discrimination, in relation to the use of high-risk AI systems referred to in Annex III "shall have the power to request and access any documentation created or maintained under this Regulation" in accessible language and format when access to that documentation is necessary "for effectively fulfilling their mandates within the limits of their jurisdiction". The relevant public authority or body shall inform the market surveillance authority of the Member State concerned of any such request" [emphasis added]. Where the documentation (subject to confidentiality) is insufficient a reasoned request can be made to the market surveillance authority, to organise testing of the high-risk AI system through technical means.

Finally, according to Article 100.1.(g), Union institutions, bodies, offices and agencies falling within the scope of this Regulation must take into account that the European Data Protection Supervisor may impose administrative fines. In doing so, due regard shall be given to their annual budget, and any funds collected by imposition of fines shall contribute to the general budget of the Union. These fines shall not affect the effective operation of the Union institution, body, office or agency fined, as explained in paragraph 6.

Conclusion

Every technology comes with its own benefits and downsides. Hence, one must always have controls in place to avoid unwanted consequences of using it. The use of new technologies in the budgetary context should be assessed considering the specific needs of different nations with regard to the protection and development of human rights. Intelligent public budgeting should care not only about not doing harm, but also about doing good with savings for better human lives.

Transparency regarding the aimed and achieved level of digitization in every step of the budget process is critical. It should cover information about the implemented technologies and their application to specific fields. The degree of openness should be sufficient to allow experts to judge their effective impact on human rights, beyond data protection.

In cases of decentralization in the allocation and execution of funds, disparity in the adoption of technology may lead to some

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

difficulties in their correct control, so more uniform approaches should be enhanced. Now there is an opportunity to guide the behavior of public entities on a large scale as to make their financing actions more respectful of citizens' rights. In practice, technological developments in budgeting and auditing can ensure the effectiveness of conditionality —at least to the extent that compliance with the requirements for the enjoyment of public financial support may be checked in real time. This chance to receive continuous feedback can lead to easily adjust the rules.

The careful use of some tech tools can help to fulfill the mandate of Article 31.2 of the Spanish Constitution, that reads: "Public expenditure shall make an equitable allocation of public resources, and its programming and execution shall respond to the criteria of efficiency and economy". Existing gaps on the public expenditure side could be reduced doubly, for example, in the design and application of benefits. On the one hand, the legislation in force could be optimized to maximize the scope of coverage by including new public needs —that are currently unattended due to lack of resources, and to improve the intensity in the degree of protection of some rights already contemplated. On the other hand, when applying the rules, budgetary execution could avoid irregularities and corruption.

Of course, any conflicting interests must be addressed: the needs of a 'machine' should not be blindly put before the person's needs when it comes to setting priorities in the context of scarce funding. It is essential to weigh up costs and benefits of strategies for the implementation in specific cases and to provide appropriate remedies (e.g. access to affordable power as many public services will require it). For similar reasons, it seems to us somewhat risky to earmark saved funds thanks to digitization only for that purpose.

The main rules that sustain our budgetary system could be gradually revised in order to integrate in a systematic manner new options offered by the evolution of technological tools. As long as they allow a better performance in carrying out the mission entrusted to budgetary and audit authorities, the legislation should be adapted to the social reality and timing in which has to be applied. For instance, one more could expect more flexibility in the regulations affecting budget appropriations. The experience in the digitalization of tax field shows clearly that tools that may modify administrative procedures, often bring with them institutional and regulatory changes as well. The public revenue and expenditure should progress together.

In the EU, the Juncker Commission started the initiative "Collect more, spend better" in the path towards a sustainable Europe by 2030

reinforcing SDG17 on Partnerships for the Goals. As the European Commissioner Neven MIMICA explained: "Collect more focuses on the efficiency, effectiveness, fairness and transparency of the tax systems at the national and international levels. This includes closing tax gaps arising from poor tax policies and from weak tax collection and enforcement under existing policies. Spend better is about improving the effectiveness and efficiency of public spending, with a particular focus on subsidy programmes, public investment, public procurement and debt. Better management in these areas can be as effective in increasing fiscal space as receiving additional resources"[19]. Even if the competent authorities, when implementing new technological means, can spend better in social policies to fulfil some human rights, policy makers and managers should bear in mind the relevance of the principle of proportionality, as these technologies in themselves, by their operation or by their result may negatively affect other human rights as well. For that reason, they should facilitate an internal and external judgment in this regard (*ex ante* rather than *ex post*).

References

Berryhill, Jamie, Kevin K. Heang, Rob Clogher & Keegan McBride. 2019. «Hello, world: Artificial Intelligence and its use in the public sector.» *OECD Working Papers on Public Governance* 36. doi.org/10.1787/726fd39d-en

Collovà, Claudio, Velina Lilyanova & Nele Lüker. 2024. «Digital public services in the National Recovery and Resilience Plans. Mid-term multilevel governance appraisal.» *Briefing European Parliamentary Research Service* PE 762.287.

European Parliament. 2024. «Briefing for the FEMM delegation to UN CSW 68 (18-22 March 2024).» *Briefing Directorate-General for Internal Policies* PE 760.544.

European Parliament. 2021. Proceedings of the workshop on «Use of big data and AI in fighting corruption and misuse of public funds - good practice, ways forward and how to integrate new technology into contemporary control framework.» *Directorate General for Internal Policies of the Union*. PE 691.722.

[19] European Commissioner Neven MIMICA, Side Event on International Support to Domestic Resource Mobilisation, Addis Abeba, 15 July 2015. http://ec.europa.eu/commission/2014-2019/mimica/announcements/side-event-collect-more-spend-better-contribution-2030-agenda-sustainable-development-new-york_en [18 October 2015]; https://www.un.org/en/development/desa/usg/statements/mr-wu/2015/09/a-contribution-to-the-2030-agenda-for-sd.html [24 November 2024]

Fedeli, Mark. 2024. «Power through fiscal year-end budgeting with AI, workflow automation.» Accessed May 22, 2024. https://www.federaltimes.com/opinions/2024/04/22/power-through-fiscal-year-end-budgeting-with-ai-workflow-automation/#:~:text=More%20federal%20agencies%20are%20harnessing,traceability%20of%20data%20across%20programs.

Fernández-Cortez, Vanessa, David Valle-Cruz & José Ramón Gil-García. 2020. «Can Artificial Intelligence help optimize the public budgeting process? Lessons about smartness and public value from the Mexican Federal Government.» *IEEE Xplore* 978-1-7281-5882-2/20.

Grau, Mª Amparo. 2023. «La utilización de la inteligencia artificial en la función de control.» *Revista Española de Control Externo* 74-75: 72-91.

Grau, Mª Amparo. 2022a. «Fiscal transformations due to AI and robotization: where do recent changes in tax administrations, procedures and legal systems lead us?». *Northwestern Journal of Technology and Intellectual Property* 19: 325-363. Accessed: May 24, 2024. https://scholarlycommons.law.northwestern.edu/njtip/vol19/iss4/1

Grau, Mª Amparo. 2022b «The Alignment of Taxation and Sustainability: Might the Digital Controls of Non-Financial Information Become a Universal Panacea?» *Review of European & Comparative Law* 3: 61.

Grau, Mª Amparo. 2020 «Los Derechos Humanos en el siglo XXI: ¿Cómo financiar su coste para salvaguardar su efiacia?» In *Los Derechos Humanos en el siglo XXI,* Tomo III: Los Derechos Humanos desde la perspectiva política y social, edited by José Antonio Pinto and Ángel Sánchez de la Torre, 175-180. Madrid: Edisofer.

Johnston, Liz. 2023a. «What State and Local Governments Should Look for in a Cloud Budgeting Software Solution.» January 19. Accessed: May 24, 2024: https://www.resourcex.net/blog/what-state-and-local-governments-should-look-for-in-a-cloud-budgeting-software-solution

Johnston, Liz. 2023b. «GFOA Budgeting Best Practices for State and Local Governments.» February 16. Accessed: May 24, 2024: https://www.resourcex.net/blog/gfoa-budgeting-best-practices-for-state-and-local-governments

Johnston, Liz. 2023c. *Step-by-Step Guide to the Public Sector Budgeting Process.* April 27. Accessed: May 24, 2024: https://www.resourcex.net/blog/step-by-step-guide-to-the-public-sector-budgeting-process

Johnston, Liz 2022b. «What Is Government Budgeting Software? 3 Basics to Know.» September 8. Accessed: May 24, 2024: https://www.resourcex.net/blog/government-budgeting-software-3-basics-to-know

Lilyanova, Velina. 2024. «Investment in artificial intelligence in the National Recovery and Resilience Plans.» *Briefing European Parliamentary Research Service*. PE 762.288.

Maciejewski, Mariusz. 2024. «Law and ICT, Policy Department for Citizens' Rights and Constitutional Affairs.» *Directorate-General for Internal Policies*. PE 762.738.

Mahmutaj, Klentiana. Forthcoming. «Artificial Intelligence, Regulation, and the Right to Development.» Thematic study by the Expert Mechanism on the Right to Development. Accessed: May 24, 2024: https://www.ohchr.org/en/documents/ongoing-studies/artificial-intelligence-regulation-and-right-development-thematic-study

Malan, Jack and Marta Dimauro. 2022. «Single Audit Approach - Root Causes of the Weaknesses in the Work of the Member States' Managing and Audit Authorities.» *European Parliament STUDY*. PE 732.267.

Martín López, Jorge. 2023. Inteligencia artificial y comprobación tributaria: transparencia y no discriminación, Pamplona: Aranzadi.

Navarra, Cecilia, Aleksandra Heflich and Meenaakshi Fernandes. 2024. «Improving EU action to end poverty in developing countries. Cost of non-Europe report.» *European Parliament STUDY.* PE 747.425.

Preut, Anna, Jan-Philip Kopka and Uwe Clausen. 2021. «Digital twins for the circular economy.» *Sustainability* 13 (18): 10467. Doi 10.3390/su131810467

Rampton, James et al. 2024. «The future of digitalisation in budgetary control.» *Directorate-General for Internal Policies*. PE 759.623.

Resource X. 2023. «Artificial Intelligence for budgeting: maximizing resources with AI modeling.» November 14. Accessed: July 12, 2024. https://www.resourcex.net/blog/artificial-intelligence-for-budgeting-maximizing-resources-with-ai-modeling

United Nations. 2022. *Revolutionizing public financial management with cutting-edge AI-powered tools.* Accessed: may 13, 2024. https://www.unescwa.org/AI-budgeting

United Nations. 2023. «Financing for development gateway: Financing is about transforming lives not just economies.». Accessed: August 21, 2024: https://www.unescwa.org/sites/default/files/inline-files/23-00158-Financing-for-Development-Gateway-Web.pdf

Valle-Cruz, David, Vanessa Fernández-Cortez and Ramón Gil-García. 2022. «From e-budgeting to smart budgeting: Exploring the potential of artificial intelligence in government decision-making for resource allocation.» *Government Information Quarterly* 39: 101644.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

Annex

Table 2.

Benefits and limitations of using risk-scoring tools to detect irregularities

Benefits	Limitations
Saving time. Because risk-scoring tools can check almost infinite amounts of data for patterns or indicators of risk they save auditors time. **Allowing a 100% check.** Risk-scoring tools may allow auditors to check not just a random sample, but all cases. **Minimising human errors.** Automating manual searches reduces human errors, and increases the chances of finding any cases of fraud and corruption. **Deterrent effect.** The increased transparency offered, in particular, by public risk-scoring tools could have a deterrent effect.	**Time and cost to develop the system.** Developing indicators, finding appropriate data, and developing risk-scoring tools to mine that data is a time and resource-intensive process. **May not capture new indicators of fraud.** Most risk-scoring tools are based on the indicators auditors defined based on their own experience. New ways to commit fraud may not be detected. This may be an issue in fast-paced environments where types of irregularities change over time. **False positives.** Not every case that is 'flagged' is fraudulent. Working with risk scores requires a level of digital literacy. While risk scores are designed to point auditors to cases to examine in more detail there is a danger that auditors may automatically see them as "fraudulent".

Source: Rampton et al. (2024).

Table 3.

Benefits and limitations of the use of digital platforms for collaboration between institutions in budgetary control

Benefits	Limitations
Enhanced efficiency: digital platforms streamline workflow by allowing effective communication and information sharing between different teams and institutions. This can increase the efficiency of an investigation and prosecution activities, where timely provision of good quality data plays a vital role. The platforms can also enable automation of manual tasks, which saves time and reduces errors. **Real-time collaboration** and interactions among audit teams can be facilitated by digital platforms, regardless of their physical location. Audit team members can also feel more motivated as they actively participate in decision-making and problem solving, which can further encourage innovation. **Single repositories of data** facilitate smoother audits as all parties have access to the same information. This can reduce confusion caused by multiple versions of documents. **Document version control and tracing** can be enabled by digital platforms, including an audit trail of revisions. **Progress tracking:** digital platforms facilitate tracking progress, milestones and tasks within budgetary control. Continuous monitoring of progress can help in ensuring timely completion of the audits.	**Cybersecurity and data privacy concerns:** the uptake in the use of digital platforms raises concerns about data protection and vulnerability to cyber-attacks. Ensuring robust cybersecurity measures is crucial. Furthermore, clear data ownership arrangements between financial institutions should be made before starting the audit process. **Adequate skills and training are necessary** for successful implementation of digital platforms. **Unforeseen technological limitations:** the full range of capabilities and limitations of digital platforms and other digital technologies is not yet fully understood. Moreover. advancements in technology might not always keep pace with the needs of the audit process. **Interoperability challenges** might arise if financial institutions use different digital platforms. Collaboration and seamless data exchange between multiple platforms might be hindered, especially where data formats are inconsistent. Multilingual data collected from various sources must be standardised so that it can be integrated into a single platform. This requires advanced translation software that can handle technical and financial terminology accurately. **Legal and regulatory compliance** with data protection laws and audit standards needs to be ensured; institutions collaborating across borders could face legal complexities. **Resistance to change:** institutions may be accustomed to traditional audit methods and resist implementing digital platforms.

Source: Rampton et al. (2024).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 173-201
https://doi.org/10.18543/djhr.3194 • http://djhr.revistas.deusto.es/

Table 4.

Benefits and limitations of the use of robotic process automation in budgetary control

Benefits	Limitations
Improved operational efficiency and resource optimisation by reducing the time and effort to complete repetitive tasks, thus allowing the audit teams, to focus on more complex issues related to audit findings. **Reduction of costs** in the long term after the initial investment to implement the technology by reducing the need for human labour or enabling staff to focus on higher-value or more complex tasks. In addition, RPA software can perform the automated tasks round the clock. **Improved compliance and data security:** automation can ensure that processes are carried out in compliance with current regulations and in a consistent manner. In addition, the risk of data breaches or unauthorised access can be reduced through automation of data encryption and access control. **Boosted accuracy** of data entry and processing by reducing the risk of errors. RPA can provide an audit trail, which makes it easier to monitor progress and resolve issues more quickly. **Easy integration** with existing legacy systems within an organization, as well as relatively straightforward implementation process. Moreover, RPA does not necessarily require a developer to configure, which makes it ideal in cases where resources are too scarce to develop deep integrations.	**Shift In organisational culture:** as RPA deployment requires a focus on more complex tasks, the adaptability of staff Is an important factor for successful outcomes in automation and digital transformation projects. Teams can be trained to adapt to the shifts in priorities. **Unable to automate more complex tasks** that require advanced decision-making as only processes with well-defined rules can be automated. **Can be difficult to scale up:**·the limited ability to handle large volumes of data may hinder RPA adoption. **Unable to learn from past experiences** and needs human intervention to learn from data and to adapt to new situations.

Source: Rampton et al. (2024).

Ética, desafíos y riesgos del acceso a la justicia algorítmica

Ethics, challenges and risks in access to algorithmic justice

José Carlos Fernández Rozas
Universidad Complutense de Madrid, España
jcfernan@der.ucm.es
ORCiD: https://orcid.org/0000-0001-8443-8488

https://doi.org/10.18543/djhr.3195

Fecha de recepción: 19.05.2024
Fecha de aceptación: 09.08.2024
Fecha de publicación en línea: diciembre de 2024

Cómo citar / Citation: Fernández Rozas, José Carlos. 2024. «Ética, desafíos y riesgos del acceso a la justicia algorítmica». *Deusto Journal of Human Rights*, n. 14: 203-235. https://doi.org/10.18543/djhr.3195

Sumario: 1. Avances de la tecnología y de la IA en la administración de la Justicia. 2. Transformación digital en el ámbito jurídico. 2.1. Nuevas aplicaciones y modelos de razonamiento. 2.2. Nuevas herramientas de tecnología jurídica. 3. Aplicaciones. 3.1. Diversidad de funciones. 3.2. Empleo por la judicatura. 3.3. Transformación de la abogacía. 4. Inteligencia artificial y acceso a la justicia. 4.1. Contradicciones en un Estado de Derecho. 4.2. Componente ético. 5. Proyección: utilidades y contraindicaciones. 5.1. Utilidades. 5.2. Presencia de sesgos. 6. Riesgos y salvaguardias. 6.1. Implementación de controles y supervisión adecuados. 6.2. Un ejemplo de la práctica: el asunto State v. Loomis. 7. Hacia una justicia algorítmica. 7.1. Metamorfosis del ámbito jurídico. 7.2. Desafíos y preocupaciones. 7.3. Necesidad de un marco regulatorio. Bibliografía.

Resumen: Socialmente, la IA suscita preocupaciones sobre la privacidad de las partes en conflicto, así como interrogantes sobre la transparencia, precisión y confiabilidad de los algoritmos utilizados en los procesos judiciales. El objetivo de este artículo es examinar los avances de la IA y sus repercusiones en los derechos humanos desde las perspectivas social y jurídica, proporcionando elementos para abordar los eventuales retos del aumento del uso de sus diferentes aplicaciones. Partiendo de sus inmensas posibilidades se sostiene que las tales aplicaciones de IA pueden ayudar a la sociedad, pero llamando la atención de las diferentes cuestiones jurídicas y la complejidad asociada derivadas en la profesión legal, analizando sus ventajas, riesgos y futuras

perspectivas profesionales. Se anticipa un cambio en las tareas profesionales, donde las actividades repetitivas serán menos valoradas, mientras que la consultoría y el asesoramiento jurídico adquirirán mayor importancia, exigiendo de los conocimientos técnico–jurídicos adaptados al desarrollo tecnológico. En términos de ética profesional, los operadores jurídicos deben entender las capacidades y riesgos de la IA, teniendo en cuenta que las actuales normas deontológicas ser actualizadas para abordar sus particularidades.

Palabras clave: Inteligencia artificial, transformación digital, tecnología jurídica, acceso a la justicia, componente ético, justicia algorítmica.

Summary: Socially, AI raises concerns about the privacy of conflicting parties, as well as questions about the transparency, accuracy and reliability of algorithms used in judicial processes. The aim of this article is to examine AI developments and their impact on human rights from social and legal perspectives, providing elements to address the eventual challenges of the increased use of its different applications. Starting from its immense possibilities it is argued that such AI applications can help society but drawing attention to the different legal issues and the associated complexity derived in the legal profession, analysing their advantages, risks, and future professional perspectives. A change in professional tasks is anticipated, where repetitive activities will be less valued, while consultancy and legal advice will become more important, requiring technical-legal knowledge adapted to technological development. In terms of professional ethics, legal operators must understand the capabilities and risks of AI, bearing in mind that current ethical rules must be updated to address its particularities.

Keywords: Artificial intelligence, digital transformation, legal technology, access to justice, ethical component, algorithmic justice.

1. **Avances de la tecnología y de la IA en la administración de la Justicia**

El rápido desarrollo de la inteligencia artificial (IA) ha transformado nuestro mundo, pero también plantea retos en la protección de los derechos humanos, lo que implica consecuencias relevantes en su protección. El avance de la tecnología y la IA en la vida diaria presenta oportunidades para mejorar la eficiencia en la administración de justicia, especialmente en la lucha contra el fraude y las irregularidades, y para que los profesionales del Derecho presten servicios jurídicos de manera más eficiente, si bien los denominados sistemas inteligentes suscitan interrogantes sobre derechos y responsabilidades individuales, así como sobre el impacto, la gobernanza y la ética, requiriendo una evaluación en términos de derechos humanos, libertades fundamentales y ética. La integración de la IA puede asistir a los profesionales del Derecho mediante *software* de gestión de casos que analiza y sugiere precedentes relevantes, aumentando la eficiencia y reduciendo costos de litigios a medio plazo, siendo esta función crucial para optimizar operaciones energéticas y desarrollar un marco político alineado con la economía circular, promoviendo la sostenibilidad. Sin embargo, plantea preocupaciones acerca de la privacidad, transparencia y confiabilidad de los algoritmos en procesos judiciales.

Aunque su aplicación en tribunales aún es experimental debido a las diferencias entre el razonamiento judicial humano y el de una máquina, en los próximos años los algoritmos predictivos aumentarán su valor añadido en el ámbito del Derecho. Los jueces deben considerar factores extrajurídicos y mantener la coherencia en la jurisprudencia, lo cual plantea dudas sobre la capacidad de la IA para prever resultados judiciales habida cuenta que la IA no está lo suficientemente avanzada para proporcionar un valor añadido significativo a los jueces. Por tanto, se requiere un marco jurídico adecuado que respete los derechos fundamentales y permita una aplicación ágil y colaborativa de la IA, protegiendo a los ciudadanos sin frenar la innovación (Fernández Rozas 2024).

Siendo indiscutida la utilidad de la IA en diversos aspectos de la justicia, su empleo limitado a predicciones o modelos matemáticos no deja de suscitar desafíos legales en áreas como la protección de datos, igualdad y transparencia, especialmente en la justicia penal (Završnik 2020, 567–583).

Concebida la IA como la inteligencia que emana de las máquinas, por oposición a la inteligencia natural de los seres humanos y los animales, se entiende como un sistema o aplicación que muestra un

comportamiento inteligente analizando su entorno y actuando con cierto grado de autonomía para alcanzar objetivos específicos. Sentado esto, deben de tenerse en cuenta una serie de matizaciones previas:

a) La definición de IA aún carece de uniformidad en el ámbito académico y legal, pero provisionalmente se identifica con la capacidad de una máquina o programa informático para realizar tareas que requieren inteligencia humana. Tradicionalmente, se divide en dos categorías: IA fuerte o general, que busca desarrollar sistemas capaces de emular o superar la inteligencia humana, e IA débil, centrada en tareas específicas mejoradas mediante el aprendizaje automático. Mientras la IA débil simula el comportamiento humano inteligente, la IA fuerte posee esta cualidad realmente.
b) La IA es una tecnología interdisciplinaria sin un propósito fijo, utilizada en diversos contextos, y su terminología es polémica debido a múltiples modelos que intentan definirla. Centrados algunos de ellos en el pensamiento y razonamiento y otros en el comportamiento, ambos pueden medir su éxito en términos de coherencia con el rendimiento humano o con una medida ideal de "racionalidad", dando lugar a cuatro grupos de modelos: "pensamiento humano", "pensamiento racional", "acción humana" y "acción racional".
c) La "justicia algorítmica" investiga cómo las instituciones gubernamentales integran la inteligencia artificial, algoritmos y tecnologías de aprendizaje automático en su toma de decisiones, y cómo esto afecta los derechos individuales y la justicia social. Su objetivo es desarrollar y utilizar aplicaciones algorítmicas que mejoren la equidad y la eficacia sin perder responsabilidad y transparencia. Esto implica considerar el impacto social, no solo la precisión técnica, y fomentar la participación ciudadana para incluir diversas perspectivas en el diseño tecnológico. Una IA fiable puede abordar desafíos sociales clave como el envejecimiento de la población, la desigualdad social y la contaminación ambiental, siendo una herramienta valiosa para mitigarlos.

Los términos de "Inteligencia Artificial" y de "aprendizaje automático" suelen utilizarse indistintamente. El aprendizaje automático se ocupa de los algoritmos (secuencias de instrucciones utilizadas para resolver un problema) de aprendizaje que, tras el análisis de gran cantidad de información (incluyendo legislación, jurisprudencia y doctrina jurídica), ofrecen resultados altamente relevantes en una

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

fracción de tiempo y toman decisiones basadas en la experiencia. Dicho aprendizaje sirve como término genérico para todos los tipos de generación de conocimiento a partir de datos, en el que un sistema aprende patrones y reglas analizando la información existente para aplicarlos a conjuntos de datos previamente desconocidos sin necesidad de programación explícita. Los algoritmos avanzados analizarán grandes cantidades de información, incluyendo jurisprudencia, estatutos y artículos legales, ofreciendo resultados altamente relevantes en una fracción de tiempo.

En el aprendizaje automático, los algoritmos aprenden de forma autónoma y mejoran continuamente con grandes cantidades de datos de entrenamiento. Existen tres tipos principales de aprendizaje: a) El algoritmo se entrena con datos completos proporcionados por humanos, buscando patrones explicativos que aplica a nuevos datos, como reconocer precios de coches basándose en características del equipamiento. b) El algoritmo identifica patrones desconocidos de forma autónoma, como detectar a personas propensas a creer y difundir mensajes falsos en redes sociales. c) El algoritmo prueba soluciones mediante ensayo y error, optimizando sus acciones a través de la interacción con el entorno. Este método se utiliza cuando hay pocos datos disponibles y no se puede determinar un resultado perfecto.

En el tratamiento de los datos legales, los analistas deben tener un profundo conocimiento del contexto y lenguaje jurídico. Idealmente, estos analistas deberían ser juristas o tener formación legal para asegurar resultados precisos y relevantes. La colaboración entre juristas y analistas de datos favorece la efectividad del etiquetado automático y otros análisis legales garantizando con ello precisión y coherencia, pues la combinación de tecnología y experiencia legal mejora la eficiencia y precisión en el análisis de sentencias y documentos legales, asegurando la integridad e imparcialidad del proceso.

2. **Trasformación digital en el ámbito jurídico**

2.1. *Nuevas aplicaciones y modelos de razonamiento*

El Derecho está siendo impactado por las nuevas tecnologías, afectando varias áreas del trabajo de los abogados, incluyendo tareas que tradicionalmente requerían juicio humano experto, como la predicción de resultados judiciales. Estas herramientas presentan retos y oportunidades: a corto plazo, se espera una mayor transparencia

jurídica, una resolución de litigios más eficiente, un mejor acceso a la justicia y desafíos para la estructura tradicional de los bufetes de abogados. Los abogados podrán trabajar de manera más eficiente, ampliar sus áreas de especialización y aportar más valor a sus clientes, transformando así su forma de trabajo y resolución de litigios. A largo plazo, el impacto de la IA en el Derecho es difícil de predecir, pero tiene el potencial de redefinir el panorama jurídico, alterando los métodos de trabajo y la estructura de las instituciones jurídicas, y mejorando el análisis y procesamiento de grandes volúmenes de datos legales.

La creciente capacidad de las herramientas de IA para realizar tareas complejas, que antes requerían habilidades humanas avanzadas, puede llevar a una reevaluación del papel del abogado tradicional que deberá adaptarse y aprender a colaborar con estas tecnologías para maximizar su potencial y evitar quedar obsoletos. Esta adaptación implica tanto una actualización técnica como un cambio en la mentalidad y cultura profesional del Derecho. Las tecnologías avanzadas ofrecen la promesa de mejorar significativamente la eficiencia y efectividad de los servicios jurídicos, planteando al mismo tiempo preguntas sobre el futuro de la profesión y cómo los abogados deberán evolucionar para seguir siendo relevantes en un mundo digitalizado, facilitando la integración de estas tecnologías la recuperación de información legal y habilita la computación cognitiva, promoviendo una colaboración estrecha entre humanos y sistemas informáticos. Esto promete mejorar la eficiencia y precisión en la toma de decisiones legales, así como una mayor comprensión y transparencia en el proceso (Ashley 2017, 38–72).

En el contexto de la comprensión del acto suministrado a partir de la IA destacan dos enfoques principales para la utilización de algoritmos. El primero, consiste en imitar el razonamiento jurisdiccional mediante sistemas expertos, lo que plantea la cuestión de si este proceso complejo puede formalizarse. El segundo, implica sistematizar la jurisprudencia a través de sistemas basados en *machine learning* (Ebers y Krõõt 2023) y en el procesamiento automático del lenguaje natural, lo que esboza la posibilidad de sustituir el razonamiento jurisdiccional por un enfoque estadístico basado en el reconocimiento de patrones. En ambos casos, surgen interrogantes acerca la viabilidad y la efectividad de estos métodos. Por un lado, la complejidad al acto de juzgar, que implica la creación de Derecho yendo más allá de meros enunciados normativos, dificulta su formalización mediante IA simbólica. Por otro lado, las limitaciones jurídicas y lógicas que enfrentan al juez esbozan dudas sobre la posibilidad de sistematizar las

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

decisiones jurídicas. Más. Al margen de estas incertidumbres, es evidente que hay una categoría de litigios cuyas soluciones son previsibles, al ofrecer las normas que los regulan un reducido margen interpretativo, lo que facilita su sistematización.

Las limitaciones técnicas de los algoritmos, como la presencia de sesgos y errores de correlación, y su incapacidad para reflexionar globalmente sobre un caso, cuestionan la viabilidad de las aplicaciones de IA para resolver litigios de forma autónoma. El sesgo es inherente a los datos, lo que puede llevar a resultados erróneos en los modelos entrenados, pues es susceptible de enmascararse, convirtiéndose en discriminación, esto es, un trato perjudicial basado en atributos sensibles, lo cual indica que la implementación inmediata de programas autónomos para la resolución de litigios podría no ser factible en ciertas situaciones.

El debate sobre la digitalización del Derecho requiere un estudio de los fundamentos del Derecho en todos sus ámbitos, involucrando disciplinas como la ciencia de datos, estadística, informática, lingüística y filosofía. Mejorar el acceso a la justicia es una misión creciente para bufetes de abogados y proveedores de servicios jurídicos (Castillejo 2022, 55–90). La centralidad del usuario impulsa la digitalización de procesos con *software* y datos, optimizando las experiencias de asesoramiento. Sin embargo, garantizar la igualdad de acceso a la justicia y asegurarla a largo plazo es también una tarea fundamental del Poder Judicial como garante del Estado de Derecho.

Están evolucionando rápidamente las aplicaciones de IA en el asesoramiento jurídico, especialmente en el ámbito de la justicia predictiva, o justicia algorítmica, que busca ofrecer soluciones confiables para problemas legales mediante el análisis estadístico de precedentes judiciales (Ashley 2017, 38–72). Sin embargo, la falta de transparencia en los algoritmos utilizados suscita preocupaciones significativas, ya que dificulta el control posterior sobre los resultados predictivos. Antes de plantear cuánta transparencia deben exigir las leyes a los algoritmos, es fundamental evaluar si las explicaciones que quienes programan pueden proporcionar son lo suficientemente comprensibles para las personas destinatarias potenciales como técnicamente factibles para quienes las producen (Mercan y Seiçuk 2024, 131–144). La exigencia de transparencia implica la necesidad de revelar los sesgos privativos de los algoritmos, pero esta revelación debe traducirse en explicaciones que sean efectivas en contextos jurídicos específicos, con el objeto de proporcionar una comprensión clara y accesible de cómo esos sesgos afectan a las decisiones algorítmicas. Esto da lugar a un sustancial desafío tanto para los

diseñadores de algoritmos como para los responsables de la formulación de políticas, ya que deben encontrar un equilibrio entre la complejidad técnica y la necesidad de comprensión y aplicación práctica en el ámbito legal (Buitten n.d., 41–59).

La extensión de la IA en la justicia predictiva podría limitar la innovación debido a su enfoque en datos preexistentes, impidiendo soluciones nuevas, siendo por ello vital evaluar los sesgos en los datos y la automatización de aspectos del litigio, como el cálculo de intereses o indemnizaciones. Los análisis muestran que los beneficios de la tecnología moderna a menudo se exageran y que la censura algorítmica afecta el acceso a la información, creando una asimetría de poder donde persistan prejuicios raciales, étnicos y de género y donde los operadores de algoritmos puedan alcanzar un control indeseado.

A medida que la IA se integra en la sociedad, se comprometen principios como accesibilidad, transparencia y equidad debido a la complejidad de sus sistemas de toma de decisiones, obligando los operadores jurídicos a informar a los consumidores sobre el uso de IA en servicios legales, especialmente con mínima intervención humana, para proteger sus derechos. Con independencia de la mejora de la calidad de los servicios legales, su uso exclusivo de la IA plantea preocupaciones legales, como las restricciones del Reglamento General de Protección de Datos (RGPD) sobre decisiones automatizadas. La justicia algorítmica puede ser beneficiosa si se usa con cautela y se reconoce la necesidad de intervención humana, pero requiere de estándares técnicos y éticos, así como de la implementación de políticas de educación y capacitación para comprender su impacto.

2.2. *Nuevas herramientas de tecnología jurídica*

Auguran los avances en IA la automatización de ciertas tareas legales, lo que podría transformar el trabajo de los profesionales del Derecho, desde abogados hasta funcionarios judiciales. Si inicialmente hubo reticencia en la profesión legal, las nuevas aplicaciones tecnológicas, capaces de aumentar la productividad, reducir constes y facilitar la vida de los usuarios, están en franca etapa de expansión, transformando una determinada percepción del mundo y modificando la manera en que se realizan muchas actividades profesionales.

El emprendimiento ha evolucionado significativamente con el concepto de *startup*, influenciado por la informatización del Derecho

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

y el avance de herramientas informáticas. La integración de tecnologías basadas en IA y el surgimiento de las *legaltechs* han digitalizado el trabajo jurídico, mejorando la eficiencia y la calidad del asesoramiento legal, y facilitando el acceso a la ley. Estos cambios están mejorando los procesos jurídicos y el mercado legal en su conjunto, haciendo el acceso a la justicia más fácil, rápido y económico. Sin embargo, esta evolución también está modificando las descripciones y requisitos de los empleos jurídicos. A pesar de los avances actuales, los procedimientos judiciales siguen desarrollándose principalmente en papel, a diferencia del arbitraje (Marrow *et al.* 2020, 35-76). La nueva normativa cambiará esta situación, mejorando significativamente la eficiencia de la cooperación judicial, el acceso a la justicia para ciudadanos y empresas, y la calidad y transparencia de la justicia.

A efectos de clarificación y como categorización intelectual resulta obligada una distinción previa entre tecnología jurídica para abogados y tecnología jurídica para usuarios finales no jurídicos. Centrándonos en la primera, resultan indiscutibles sus avances en métodos de trabajo aplicables a la vida profesional cotidiana que antes no eran posibles y generan un valor añadido por el momento imprevisibles, contemplando asimismo ideas innovadoras que satisfacen directamente las necesidades jurídicas de quienes buscan justicia utilizando la tecnología, en algunos casos evitando los órganos tradicionales de la administración de justicia (Fernández Pérez 2024).

De hecho, cada vez son más las administraciones públicas y empresas del sector privado que utilizan la IA, ya que sus aplicaciones aportan mejoras de eficiencia y ahorro de costes. Un reto jurídico particular para el sector público es la transparencia y la legalidad, toda vez que las actuaciones de la administración pública están reguladas por ley. La ley garantiza la previsibilidad y la seguridad jurídica y protege a la ciudadanía de la arbitrariedad y. en un país donde los derechos de las personas son extremadamente importantes, el principio de legalidad también garantiza la legitimidad democrática. Por tanto, el uso de la IA en la administración pública debe ser transparente y, si entraña riesgos, basarse en un fundamento jurídico adecuado. Es legítimo que las administraciones públicas se esfuercen por ser más eficientes con la ayuda de la IA. No obstante, los diversos colectivos sociales deben extremar la vigilancia y no olvidar ciertos ejemplos de utilización de la IA que han conducido a resultados desastrosos: por eso es importante analizar el impacto de tales herramientas de IA antes y después de su puesta en funcionamiento.

3. **Aplicaciones**

3.1. *Diversidad de funciones*

El avance de la IA está impregnando diversas esferas de la sociedad, y el ámbito legal no es una excepción. Los sistemas de IA se están convirtiendo en herramientas cada vez más sofisticadas para asistir a los profesionales del Derecho en múltiples tareas. Desde la investigación jurídica hasta la redacción de documentos legales y la evaluación de casos, estas tecnologías prometen mejorar la eficiencia y calidad de la justicia si se utilizan con prudencia y cuidado. Entre otras funciones estas herramientas pueden:

a) Agilizar y mejorar significativamente el proceso de investigación legal. Los algoritmos de búsqueda avanzada consiguen analizar grandes volúmenes de documentos legales y jurisprudencia para encontrar precedentes relevantes y argumentos jurídicos pertinentes. Esto ahorra tiempo a los operadores jurídicos y les permite acceder a información clave de manera más rápida y eficiente.
b) Auxiliar en la redacción y revisión de documentos legales. Utilizando técnicas de procesamiento del lenguaje natural, ayudan a redactar contratos, escrituras y otros documentos legales de manera más precisa y coherente; y también pueden realizar análisis gramaticales y de coherencia para identificar posibles errores o inconsistencias en los documentos.
c) Contribuir a la evaluación de casos y la predicción de resultados judiciales. Al analizar datos históricos de casos similares, los algoritmos están en disposición de identificar patrones y tendencias que pueden ayudar a predecir el resultado de un litigio con un cierto grado de precisión. Esto puede ser especialmente útil para evaluar la viabilidad de un caso y tomar decisiones estratégicas durante el proceso legal (Chatterjee, Singhania y Sharma 2023).

La asistencia a ciudadanos a través de *chatbots* considerando plataformas como "eJusticia" es útil para agilizar procesos judiciales mediante el análisis de expedientes, previamente practicada la anonimización y seudonimización de las decisiones, pero debe considerarse como un "complemento" para los jueces, no como un "sustituto" de sus decisiones. No en vano, su utilización en la evaluación de pruebas plantea desafíos respecto a su fiabilidad y autonomía, requiriendo una reflexión profunda sobre su aplicación

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

(Nieva 2018, 70–100). Uno de ellos es la falta de capacidad explicativa en los sistemas resultantes, susceptible de minar la confianza en sus resultados, especialmente en aplicaciones críticas, a lo que debe añadirse la posibilidad de generar resultados incorrectos cuando los datos relevantes no están incluidos en el conjunto de entrenamiento, conocido en estadística como el *long tail problem*.

Otros desafíos de transparencia y fiabilidad requieren un análisis crítico. Estas tecnologías transforman funciones judiciales y servicios legales, apoyando a profesionales mediante la automatización de procesos y asistencia en decisiones, y en la redacción y revisión de contratos, investigación legal y gestión de despachos. La expansión de la IA y la robótica necesita un marco jurídico adecuado, al margen de documentos preliminares, especialmente en el ámbito del Derecho privado, como la responsabilidad civil en vehículos autónomos y robots, y la automatización de la contratación. Esto implica un nuevo enfoque en elementos del Derecho contractual como la autonomía de la voluntad, el error, el engaño, la interpretación de contratos y la buena fe, adaptando los principios legales tradicionales a la IA y la robótica.

3.2. *Empleo por la judicatura*

La innovación tecnológica en el ámbito judicial a menudo se asocia con la idea de grandes inventos, como robots e IA, individualizados con el término de "agentes", que toman decisiones con capacidades sobrehumanas y tiempos de justicia rápidos. Sin embargo, la introducción de tecnologías ya existentes en nuevos contextos y procesos puede acarrear consecuencias significativas que alteran la actividad judicial y su percepción por parte de profesionales y público en general al ofrecer la promesa de transformar la gestión de casos legales al permitir el análisis y procesamiento eficiente de grandes cantidades de datos (Re y Solow–Niedermann 2019, 252). Como resultado, la relación entre IA y el papel del juez genera un doble movimiento: por un lado, cuestiona la capacidad de la IA para resolver litigios jurídicos; y, por otro lado, suscita la interrogante sobre cómo debe entenderse el acto de juzgar cuando está influenciado por la IA.

Debido a la IA, el sistema judicial, al igual que muchas otras áreas, está experimentando cambios y, aunque la sustitución completa de los jueces por IA sigue siendo técnicamente imposible, ya se utiliza en los tribunales para apoyar o asumir determinadas actividades. Los juristas argumentan que un "juez robot" no puede ocuparse de derechos fundamentales, y que el derecho a un juicio justo debe incluir

necesariamente a jueces humanos. La IA ahora permite reforzar tareas puramente intelectuales mediante máquinas y programas orientados a la búsqueda de datos legales, predicción de decisiones y análisis estadístico de decisiones pasadas utilizándose por la judicatura como apoyo, sin influir en el contenido de sus decisiones, pues no puede reemplazar el análisis exhaustivo y la interpretación de la ley que realizan los jueces. Sin duda ofrece ventajas de independencia y objetividad, pero sigue presentando riesgos de errores y consecuencias adversas inherentes a la ausencia de comprensión humana, por eso deben ser bien recibidos los proyectos para integrar la IA en el trabajo de los jueces, reconociéndose que las reformas regulatorias pueden facilitar el acceso equitativo a la IA jurídica, fomentando la colaboración entre servicios legales e industrias tecnológicas.

3.3. *Transformación de la abogacía*

El avance tecnológico está transformando la abogacía con la aparición de nuevas aplicaciones que aprovechan el crecimiento de datos disponibles y las capacidades de procesamiento de datos. Estas aplicaciones están ampliando los recursos disponibles para los profesionales legales y facilitando la investigación. Siendo la fiabilidad de los programas de justicia predictiva aún cuestionable, es probable que mejoren con el tiempo y sean adoptados por los abogados en su práctica diaria. Sin embargo, el impacto de estos cambios en la forma de prestar servicios legales y el derecho al asesoramiento aún es incierto, ya que también generan nuevas expectativas y hábitos entre el público. La integración de la IA tiene el potencial de ser un complemento transformador y positivo tanto para los abogados como para los clientes ofreciendo un inmenso potencial para revolucionar una amplia gama de actividades jurídicas.

Estos modelos de IA, entrenados a partir de vastos conjuntos de datos, pueden llevar a cabo investigaciones jurídicas y generar contenidos coherentes y contextualmente relevantes, desde sinopsis de casos hasta cláusulas contractuales, con notable precisión y eficiencia. La tecnología está transformando la práctica legal y los sistemas judiciales al automatizar tareas como la toma de decisiones y el análisis legal, mejorando la eficiencia y rapidez en la tramitación de casos, ampliando el acceso a la justicia. Con independencia de que regular esta evolución es difícil, ofrece oportunidades para reducir costos y mejorar la eficacia. Las tecnologías actuales van desde portales web automatizados hasta algoritmos para interpretar textos legales, liberando tiempo para

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

actividades más estratégicas. Sin embargo, la introducción de la IA plantea desafíos éticos en términos de transparencia y responsabilidad, resaltando la necesidad de garantizar la precisión y equidad de estas herramientas con la participación de los profesionales del Derecho.

Los bufetes y otras entidades privadas han estado hasta ahora a la vanguardia de la experimentación de la industria legal con la IA generativa con el consiguiente aumento de la productividad, potenciando a la vez su creatividad y eficiencia y generando a la vez nuevas ideas y perspectivas actuando como un asistente que sugiere soluciones y libera a los abogados de tareas repetitivas. Es cierto que la eficiencia puede mejorarse con IA, pero la creatividad es una cualidad humana que difícilmente puede ser reemplazada, enfrentando a las firmas de abogados al dilema de cómo adoptar la IA pues, si bien esta tecnología aumenta la eficiencia, en este modelo de estructura organizativa no se acostumbra a valorar la eficiencia sino el esfuerzo y la cantidad de trabajo realizado por los abogados, a lo que cabe añadir que la creatividad de los abogados no es incentivada tanto como el trabajo rutinario y detallado. Así, mientras que la IA puede transformar la práctica legal aumentando la productividad y creatividad individual, podría alterar los fundamentos económicos de los bufetes que dependen del trabajo intensivo en horas. Sea como fuere, la IA obligará a los bufetes a replantear su modelo de gestión, alejándose de la facturación por hora hacia un enfoque centrado en la productividad y creatividad de sus abogados lo que exigirá nuevas estructuras organizativas y modelos de valor que se adapten mejor a un entorno post–IA. La implantación de la IA augura, pues, una transformación profunda del sector legal, centrada en la eficiencia y creatividad, que podría beneficiar tanto a los abogados como a los clientes.

En el contexto actual de creciente digitalización, los abogados deben adquirir un sólido conocimiento sobre medidas de protección de datos para salvaguardar la información confidencial de sus clientes y proteger sus sistemas contra amenazas cibernéticas. Esto incluye la instalación de cortafuegos vigorosos y el uso de tecnologías para fortalecer contraseñas.

4. **Inteligencia artificial y acceso a la justicia**

4.1. *Contradicciones en un Estado de Derecho*

El acceso a la justicia es un principio fundamental del Derecho procesal civil para garantizar la igualdad y la protección de los derechos

humanos. Este acceso equilibra el campo de juego, asegurando que todos puedan enfrentar situaciones de vulnerabilidad y desigualdad, convirtiéndose en un pilar ético y moral de un Estado de Derecho democrático. Sin embargo, enfrenta obstáculos como la falta de información, los costos legales y la lentitud de los procesos judiciales. La IA podría ayudar a superar estas barreras, mejorando así el acceso a la justicia y la tutela judicial efectiva.

La efectiva realización del acceso a la justicia requiere un enfoque proporcional que considere cuidadosamente las posibilidades y desafíos específicos de cada caso. La IA ha sido anunciada por su potencial para ayudar a cerrar la brecha en el acceso a la justicia y puede aumentar la eficiencia, democratizar el acceso a la información jurídica y ayudar a los consumidores a resolver sus propios problemas legales o ponerlos en contacto con profesionales autorizados que puedan hacerlo. Pero se corre el riesgo que una mayor dependencia de la IA conduzca a uno o más sistemas de dos niveles: los menos favorecidos podrían quedar atrapados con una asistencia inferior basada en la IA y únicamente las grandes y costosas firmas de abogados podrían ser capaces de aprovechar eficazmente los beneficios que suministra la IA; o el impacto de la IA podría no alterar el *statu quo* en el que sólo algunos pueden permitirse cualquier tipo de asistencia jurídica.

Los estudios jurídicos y las nuevas tecnologías muestran una contradicción evidente: mientras la tecnología avanza rápidamente hacia la internacionalización y globalización, el Derecho reacciona lentamente y permanece confinado a las fronteras nacionales. Incluso se llega a considerar a la IA como una amenaza al Estado de Derecho al introducirse en la toma de decisiones y, en algunos casos, reemplazar a los responsables humanos, como en la asistencia a jueces. La opacidad de la IA, debido a su complejidad, genera preocupaciones sobre su protección legal, transparencia, reputación, imparcialidad en la justicia y la credibilidad de los principios del Estado de Derecho.

El Estado de Derecho exige que las leyes sean accesibles y previsibles para los ciudadanos, promoviendo la publicación y la inteligibilidad de las normativas, socavando la complejidad de la IA estos atributos, haciendo que sus aportaciones sean menos comprensibles y transparentes. La IA realiza cálculos matemáticos complejos que superan la comprensión humana, lo que dificulta explicar sus procesos y resultados de manera clara con los consiguientes desafíos para la transparencia y comprensión al producirse una contraposición entre el "lenguaje natural" y el lenguaje matemático y estadístico, a menudo oculto en una "caja negra"

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

(Greenstein 2022, 291–323). Al basarse el Estado de Derecho en un lenguaje natural, muchos procesos jurídicos pueden automatizarse con éxito. Sin embargo, a medida que la gobernanza se digitaliza, la comprensión ciudadana puede disminuir, especialmente con sistemas opacos incluso para sus creadores.

La complejidad de los algoritmos aumenta el potencial de error, subrayando la importancia del derecho a reclamar decisiones críticas tomadas por máquinas autónomas. Cuando la IA toma decisiones sin transparencia, desafía el Estado de Derecho. Si la automatización puede mejorar la eficiencia de los servicios públicos, por el contra, la opacidad tecnológica es susceptible de socavar la confianza en su uso, planteando si el futuro será un Estado de Derecho algorítmico o basado en principios humanos. La tecnología no es directamente responsable de la erosión del Estado de Derecho, pero puede perturbar el equilibrio entre los intereses en conflicto que se logra a través del derecho tradicional. En la UE el RGPD es un ejemplo de cómo el Derecho tradicional busca equilibrar los derechos de propiedad intelectual y los derechos de privacidad en el contexto de la IA.

Siendo consustancial al Estado de Derecho la capacidad de impugnar decisiones, la opacidad de la IA y las barreras legales de la propiedad intelectual debilitan este derecho, haciendo necesario integrar la impugnabilidad en el diseño de sistemas IA. La falta de transparencia dificulta identificar e impugnar decisiones automatizadas fuera del ámbito judicial, dejando a las personas sin conocimiento de decisiones que les afectan, siendo menester establecer un mecanismo de notificación que informe a las personas cuando han sido objeto de una decisión tomada por IA, implementando un "derecho a saber" que permita a los individuos conocer y comprender dichas decisiones.

El debido proceso, fundamental en cualquier procedimiento judicial, garantiza la protección contra abusos de poder y arbitrariedades, proporcionando a los ciudadanos instrumentos para defenderse, y se configura como un derecho humano fundamental que busca asegurar la igualdad de derechos para todos los individuos (Añón 2018, 31). El derecho de acceso a un tribunal de justicia, también conocido como derecho a la tutela judicial efectiva, está consagrado en el en el art. 8 Declaración Universal de Derechos Humanos y en el art. 471 de la Carta de los Derechos Fundamentales de la Unión Europea, así como en el art. 6 del Convenio Europeo de Derechos Humanos (CEDH) (Leslie *et al*, 2021, 3). y en el art. 24.1 de la Constitución Española, e implica no solo la existencia de recursos legales, sino también su efectividad y eficacia. Bien entendido que el

ejercicio de este derecho, al estar sujeto a limitaciones implícitas, debe para garantizar que las personas cuyos derechos y libertades han sido vulnerados tengan la posibilidad de recurrir a la justicia (Gómez Colomer 2022, 275–276).

En este contexto, es importante que la utilización de la IA para resolver controversias en línea, a través de los procedimientos de Resolución de Disputas en Línea (ODR), no menoscabe el derecho de acceso a un tribunal (Scherer 2019, 539–573). En 2015, la Asamblea Parlamentaria del Consejo de Europa destacó la importancia de garantizar que las partes involucradas en estos procedimientos de ODR conserven el derecho a recurrir a un proceso de apelación judicial que cumpla con los requisitos de un tribunal justo, de acuerdo con el art. 6 CEDH. En todo el mundo, el acceso a la justicia sigue siendo un desafío importante desde hace décadas.

4.2. *Componente ético*

El creciente uso y relevancia de la IA en todos los ámbitos de la vida y de los negocios ha desencadenado en los últimos años un amplio debate público sobre sus consecuencias. Aspectos éticos como la falta de transparencia en las decisiones respaldadas por IA, la discriminación por algoritmos y la preocupación por la protección de datos están en el centro de este debate (Esparza 2022, 181). A pesar de los numerosos ejemplos individuales, en la actualidad sólo existen unos pocos estudios que examinen de forma holística el impacto real de la IA sobre los derechos fundamentales. Independientemente de las diferencias entre sectores, la IA ofrece en general oportunidades para las empresas, la economía en su conjunto y el sector público. Dadas las oportunidades económicas, resulta obligado analizar su impacto en los derechos fundamentales. En particular, cuando la IA sustituye (en lugar de simplemente apoyar) la toma de decisiones humana, pueden surgir problemas éticos y de derechos fundamentales.

Es susceptible la IA de impactar en diversos derechos fundamentales contemplados en la Carta de los Derechos Fundamentales de la Unión Europea (CDFUE): a) En el ámbito de la dignidad, puede tener efectos tanto positivos como negativos, especialmente en la asistencia sanitaria; b) en cuanto a la libertad, su empleo puede afectar especialmente al derecho a la vida privada y a la protección de los datos personales, dado el uso de grandes cantidades de datos; c) también puede influir en otros derechos fundamentales, como la igualdad y la solidaridad; por ejemplo,

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

los algoritmos pueden generar sesgos que afecten al derecho a la no discriminación, como se ha visto en casos de decisiones judiciales sesgadas por algoritmos; d) en orden a la solidaridad, se debate si la IA podría aumentar el desempleo, al margen de la existencia de aplicaciones positivas, como en la protección del medio ambiente; e) en relación con los derechos de la ciudadanía, la IA podría mejorar la eficiencia de la administración pública, pero también plantea riesgos para la integridad de las elecciones al facilitar la difusión de información falsa para influir en los votantes; f) por último, la Carta establece que toda persona tiene derecho a la tutela judicial efectiva, a un juez imparcial y a la defensa; de esta suerte, la IA utilizada en el sistema judicial conduce a la problemática relativa al derecho a un juicio justo (*v.gr*., si tales sistemas contienen sesgos).

Por su parte, el Consejo de Europa, en el seno de la Comisión para la eficiencia de la Justicia (CEPEJ), aprobó en 2018 la Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno donde se recogen cinco principios éticos que deben informar el uso de IA por parte de un juzgado o tribunal, pero también cualquier operador jurídico que trabaje en su entorno.

La llegada de la IA ha abierto numerosas posibilidades para mejorar la eficiencia y calidad en la administración de justicia. Si se utiliza de manera responsable y ética, puede incrementar la eficiencia, accesibilidad y calidad de la justicia. La intersección entre ética y Derecho de la IA busca buscan establecer normas y valores que rijan la conducta humana y los sistemas sociotécnicos habilitados por la IA. Los derechos humanos y los valores sociales son elementos centrales en ambos campos, proporcionando un marco normativo para políticas y regulaciones que promuevan el uso ético y responsable de la IA (Sartor 2020, 708–719). Esta convergencia fomenta un enfoque holístico que aborda consideraciones éticas y legales, contribuyendo al desarrollo equilibrado y sostenible de la IA en la sociedad. En el contexto actual, la IA es una herramienta valiosa para mejorar el acceso a la justicia y fortalecer los derechos de defensa, siempre que se maneje con precaución y se promueva la transparencia y alfabetización digital en su uso. La IA puede proporcionar a las partes en un juicio acceso a información relevante para comprender mejor las decisiones judiciales y presentar argumentos fundamentados. No obstante, el aseguramiento de estos beneficios requiere la alfabetización digital entre los participantes del sistema judicial y el establecimiento de mecanismos de transparencia que permitan entender y cuestionar el funcionamiento de los algoritmos y las decisiones generadas.

5. **Proyección: utilidades y contraindicaciones**

5.1. *Utilidades*

El acceso a la justicia y la digitalización del sistema judicial constituyen, en definitiva, dos caras de la misma moneda y en la medida en la cual el sistema judicial puede digitalizarse de forma integral y fundamental, la aceptación y la eficiencia del sistema jurídico aumentará, indudablemente, de forma masiva. Cabe cuestionarse, sin embargo, la afectación que la IA podría ejercer en el derecho a un juicio justo; esto es, si una gran firma, al disponer de mayores medios que un pequeño despacho, podría unir datos públicos y privados con mayor facilidad y fidelidad, obteniendo así un perfilado muy concreto del juez que deba resolver un asunto y generando un desequilibrio en el caso. También se cuestiona si el derecho a un recurso efectivo puede verse comprometido por las aplicaciones de la IA. En principio, todos los derechos humanos mencionados exigen transparencia, trazabilidad (capacidad de rastrear y seguir el recorrido de un producto desde su origen hasta su destino final) y "explicabilidad" de las aplicaciones de la IA, ya que es la única manera de garantizar el enjuiciamiento efectivo de una violación de tal derecho humano puede garantizarse. Esto se debe a que parte del derecho a un recurso judicial efectivo es la garantía de una decisión motivada e individualizada.

Las obligaciones de transparencia pueden incluir, entre otras cosas obligaciones de notificación u obligaciones de divulgación. La Sentencia del TEDH de 4 de diciembre de 2015, en el asunto *Zakharov c. Rusia* consideró que, en el caso de algoritmos automatizados utilizados para la prevención de delitos, como la vigilancia secreta del tráfico móvil, era necesaria la notificación para garantizar la eficacia de los recursos legales. Por último, también debe ser posible una indemnización adecuada, pero no basada puramente en términos de pérdidas económicas, sino que deba tener en cuenta adecuadamente el daño por la violación de un concreto derecho humano.

La IA asiste a particulares, litigantes y jueces en la organización y enriquecimiento de la información legal, proporcionando consejos y sugerencias útiles. Por ello los jueces deben comprender el funcionamiento de la IA y sus avances para utilizarla de manera efectiva. Es ineludible, pues, que los tribunales digitalicen su información y la doten de interpretación jurídica para que la IA pueda aprovecharla al máximo. Con ello pueden alcanzarse resultados como los siguientes:

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

a) Eficiencia. La IA optimiza la eficiencia y velocidad en los procedimientos judiciales, reduciendo la carga de trabajo de los profesionales legales, permitiendo resoluciones más rápidas y minimizando errores humanos con análisis objetivos y recomendaciones basadas en datos. Esto mejora la precisión y equidad en la toma de decisiones. Automatiza tareas rutinarias, acelera los procedimientos, disminuye la carga de trabajo de jueces y abogados, y mejora la evaluación de la evidencia al analizar grandes volúmenes de pruebas. Sin embargo, es crucial no prescindir de la interpretación humana en las decisiones judiciales para evitar desafíos éticos y prácticos.
b) Duración razonable. En el ámbito judicial, un volumen significativo de trabajo depende de investigaciones y redacciones individuales, lo que genera desafíos internos en capacidad y calidad, y una crisis externa por retrasos inaceptables. El art. 6.1. CEDH enfatiza la necesidad de procedimientos judiciales de duración razonable, obligando a los Estados a organizar eficazmente el sistema para evitar demoras. La IA puede simplificar y acelerar el proceso judicial, especialmente en contextos de escasez de personal. Estudios empíricos demuestran que la IA puede analizar contratos más rápido y con mayor precisión que los abogados, sugiriendo su potencial para mejorar la eficiencia del sistema judicial. Integrar adecuadamente la IA podría ayudar a abordar estos desafíos y asegurar una justicia más ágil y efectiva.
c) Costes. En el ámbito legal, un desafío principal para acceder a la justicia es el costo del asesoramiento y representación legal. Los programas de asistencia jurídica pueden mejorar estas estrategias, toda vez que evaluación de circunstancias personales y financieras para la asistencia legal y costos judiciales suele requerir decisiones claras. Los sistemas algorítmicos, técnicamente viables, podrían ser una solución eficiente para estas evaluaciones.

Herramientas como las que se han indicado, podrían agilizar el proceso de evaluación, garantizando una asignación más justa y eficiente de los recursos disponibles para la asistencia jurídica, manifestándose en una serie de situaciones, entre las que pueden apuntarse, a título meramente ejemplificativo, las siguientes:

a) Conflictos de baja intensidad. La tecnología está demostrando ser una herramienta eficaz para la resolución rápida y económica de conflictos de baja intensidad en el ámbito legal. Plataformas

de resolución en línea, respaldadas por IA, están facilitando la solución de litigios sin necesidad de procesos judiciales costosos y prolongados, respondiendo a una demanda creciente de mecanismos simplificados. Esta tendencia no solo está cambiando la forma en que se abordan los conflictos legales, sino también transformando el propio sistema legal. Sin embargo, aún queda mucho por explorar en cuanto a los posibles usos de la IA en el ámbito jurídico, especialmente debido a la incertidumbre sobre las herramientas inteligentes disponibles y su aplicación real. La IA podría facilitar la resolución de conflictos de baja intensidad que no plantean cuestiones éticas y jurídicas complejas, lo cual es beneficioso ya que estos conflictos no requieren intervención judicial debido a su naturaleza básica y predecible. Esto podría mejorar la confianza en la institución judicial y fortalecer los principios del Estado de Derecho en democracias que enfrentan crisis de definición.

b) Personas con discapacidad. Los sistemas de IA mejoran la asistencia legal mediante herramientas de traducción automática, comprensión de documentos y recursos adaptados, lo que promueve una justicia más inclusiva para personas con discapacidad. Con financiación pública y sin fines de lucro, se podrían desarrollar aplicaciones tecnológicas legales para atender necesidades especiales, como plataformas accesibles y asesoramiento jurídico de bajo umbral. Estas aplicaciones también pueden impulsar mejoras sostenibles en el sistema legal. El papel del facilitador ha de garantizar la participación de las personas con discapacidad, proporcionando apoyo e información accesible, asegurando igualdad de oportunidades. La integración de la IA complementa el trabajo del facilitador en la justicia inclusiva.

c) Acceso igualitario a la justicia por parte de las mujeres. El acceso a la justicia, como derecho fundamental, debe garantizarse plenamente, especialmente para las mujeres, considerando las barreras y prejuicios de género arraigados en las instituciones jurídicas. La integración de una perspectiva de género en el desarrollo y uso de sistemas de IA en el ámbito judicial es esencial para corregir desigualdades y mejorar el acceso igualitario a la justicia y la IA puede ser una herramienta para revelar y abordar realidades injustas, permitiendo la denuncia de dinámicas que vulneran los derechos de las mujeres, lo que se consigue incorporando la perspectiva de género en la

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

implementación de esta tecnología para eliminar obstáculos y avanzar hacia la igualdad de género en el acceso a la justicia.

d) Personas con bajos ingresos, como las que son objeto de discriminación de género, como trabajadoras precarias, informales, desempleadas o madres solteras, la comprensión del marco legal que regula sus vidas cotidianas (Carrizo 2019, 287–310). Esto determina en gran medida su ejercicio de la ciudadanía y las consecuencias de sus decisiones. Su conocimiento del ordenamiento jurídico puede influir en aspectos como su estatus legal, acceso a beneficios ciudadanos, protección como consumidoras, derechos como inquilinas o madres, y su tratamiento justo en situaciones legales como separaciones, divorcios o la custodia de sus hijos (Ortiz de Zárate y Guevara 2021).

5.2. *Presencia de sesgos*

También puede ser considerado el uso de sistemas de IA y algorítmicos en la revisión de las circunstancias personales y económicas en su determinación para recibir asistencia legal. En el contexto de la justicia digital, los algoritmos pueden parecer neutrales, pero en realidad son susceptibles, en unos casos, de reflejar los sesgos de sus diseñadores, perpetuando la discriminación contenida en los materiales introducidos en el sistema y, en otros casos, de ser lo suficientemente opacos dificultando la averiguación o la corrección de los sesgos, lo que plantea serias preocupaciones sobre la imparcialidad del proceso judicial. Y ello sin olvidar que la falta de transparencia y motivación en las decisiones del sistema judicial digital también socava la confianza en su imparcialidad, ya que los litigantes no pueden verificar cómo se llegó a una determinada conclusión (Nieva 2018, 144–146).

Indudablemente las decisiones humanas también pueden estar sesgadas, los algoritmos tienen el potencial de amplificar y perpetuar estas desigualdades de manera sistemática, lo que añade un desafío para mantener la imparcialidad en las decisiones judiciales. Pese una apariencia de imparcialidad los algoritmos pueden generar decisiones discriminatorias, incluso si fueron diseñados con buenas intenciones, contraviniendo la prohibición de discriminación en los derechos fundamentales. Por eso, los desarrolladores deben corregir estos sesgos, lo que implica manejar definiciones filosóficas sobre la igualdad

y tomar decisiones sobre qué elementos favorecer o neutralizar. Minimizar los sesgos en la IA requiere abordar los sesgos presentes en los datos de origen y los sistémicos acumulados durante décadas. No se trata de evitar la tecnología, sino de buscar soluciones que equilibren las decisiones basadas en máquinas y las humanas, proponiendo soluciones concretas a los problemas identificados.

6. **Riesgos y salvaguardias**

6.1. *Implementación controles y supervisión adecuados*

La IA ofrece una oportunidad valiosa para mejorar la eficiencia y precisión del proceso judicial, pero su implementación requiere precauciones para evitar violaciones de derechos fundamentales. A medida que la IA se integra cada vez más en la vida cotidiana, la comunidad jurídica muestra un interés creciente al considerar que la tecnología robótica puede beneficiar el bienestar humano, pero no deben olvidarse los riesgos y desafíos asociados con su uso no regulado. Por eso, es necesario la toma en consideración de la normativa internacional y nacional para garantizar un uso responsable de la IA que proteja los derechos humanos y promueva el bienestar general. Al regular la IA, es necesario considerar su naturaleza legal y la posibilidad de otorgarle personalidad jurídica, especialmente en términos de su impacto en los individuos y los derechos humanos.

El uso de tecnologías de IA en los procedimientos jurisdiccionales ofrece oportunidades para mejorar la aplicación del derecho a la tutela judicial, pero también presenta riesgos de violación del derecho a un juicio justo. Esto incluye: a) establecer salvaguardias para un uso justo, transparente y ético de la IA, junto con políticas que aseguren la rendición de cuentas y supervisión; b) identificar y mitigar sesgos algorítmicos, protegiendo la privacidad y derechos fundamentales; y c) equilibrar la eficiencia y justicia con la protección de los derechos humanos y el Estado de Derecho. Esto es, un enfoque colaborativo que involucre a legisladores, autoridades judiciales, expertos en tecnología y defensores de derechos humanos.

La autonomía de los sistemas de IA aumenta el riesgo de usos indebidos y accidentes por instrucciones mal especificadas. Para aprovechar su potencial, es crucial que la tecnología sea accesible y diseñada según las necesidades de los usuarios, asegurando una distribución equitativa de sus beneficios evitando la manipulación por parte del agente, su desarrollador, el usuario o entidades sociales, con

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

el establecimiento de salvaguardas que garanticen transparencia, imparcialidad y equidad en el uso de la IA en el sistema judicial, la implementación de controles adecuados y la capacitación del personal judicial en su uso.

El avance de la IA ha llevado a su integración en procesos administrativos y judiciales, complementando o sustituyendo a responsables humanos en áreas como la policía predictiva y la evaluación automatizada del riesgo penal y, si bien puede mitigar sesgos humanos y mejorar la eficiencia, también presenta riesgos significativos al reducir el componente humano en decisiones gubernamentales. Por ello, es necesario desarrollar marcos normativos sólidos que garanticen transparencia, equidad y rendición de cuentas en el uso de la IA por parte de los gobiernos (Schäferling 2023).

El uso de la IA presenta tanto oportunidades como riesgos, especialmente debido a la falta de regulación específica. En áreas como el reconocimiento facial y la contratación de personal, la IA puede introducir sesgos que fomentan actitudes discriminatorias, afectando a trabajadores y exponiendo a empleadores a sanciones legales. La ausencia de un marco jurídico específico para la IA sugiere que la regulación actual no aborda adecuadamente estas discriminaciones en el proceso de contratación. Legisladores e industria deben establecer condiciones adecuadas para maximizar ventajas y minimizar riesgos. En los últimos años, han surgido diversas iniciativas no legislativas a nivel nacional e internacional y muchos agentes privados han adoptado medidas de autorregulación.

6.2. *Un ejemplo de la práctica: el asunto State v. Loomis*

El programa COMPAS es una herramienta de justicia predictiva utilizada en varios Estados de EE.UU., que ha sido objeto de debate debido a preocupaciones sobre su equidad y discriminación. Utilizando algoritmos de aprendizaje automático, evalúa el riesgo de reincidencia de los acusados para ayudar a los jueces a determinar las penas. A través de un cuestionario de 137 preguntas, el programa asigna una puntuación de peligrosidad en una escala del 1 al 10, clasificando el riesgo como bajo (verde), medio (naranja) o alto (rojo). Aunque no incluye preguntas sobre origen étnico, se ha demostrado que perpetúa prejuicios raciales indirectos basados en otros criterios, conocidos como *proxies* (Forrest 2021).

En el caso *State v. Loomis*, resuelto en 2016 por el Tribunal Supremo de Wisconsin, Eric Loomis fue condenado a prisión basándose

en una evaluación de riesgo realizada por el programa COMPAS. Loomis impugnó su condena, argumentando que el uso del COMPAS violaba sus derechos procesales porque el código fuente del software no fue revelado, impidiéndole cuestionar la validez científica de los resultados. Además, alegó que el programa utilizaba su sexo como factor en la evaluación, violando su derecho a una condena individualizada y a la igualdad de protección según la XIV Enmienda. El Tribunal Supremo de Wisconsin, al abordar por primera vez la constitucionalidad del uso de algoritmos en la imposición de penas, aceptó el uso del COMPAS, argumentando que el derecho de los acusados a las garantías procesales no se veía vulnerado por no poder acceder a una explicación del algoritmo. El Tribunal consideró que el derecho del acusado a una condena individualizada había sido violado, pero rechazó la demanda entendiendo que la precisión de las herramientas y la capacidad de los jueces para comprender su posible mal funcionamiento eran suficientes para garantizar los derechos de los acusados. Así las cosas, Loomis apeló ante el Tribunal Supremo de EE.UU., pero este decidió el 18 de septiembre de 2017 no considerar el caso, lo que confirmó la decisión del Tribunal Supremo de Wisconsin. Las razones esgrimidas por el Tribunal Supremo de Wisconsin merecen un análisis detallado por mostrar cierto respaldo hacia programas como COMPAS, afirmando que estos son solo uno de los muchos factores a considerar en la imposición de la pena. Destacó que la puntuación de riesgo del COMPAS no puede ser decisiva por sí sola, ya que eso sería contrario a la ley. En resumen, el Tribunal sostuvo que, al no poder basarse exclusivamente en los resultados del COMPAS, no se infringía el debido proceso. En este caso el Tribunal Supremo, enumeró cinco razones que justificaban la cautela: en primer lugar, el programa informático COMPAS era de propiedad privada y carecía de transparencia en su funcionamiento; en segundo lugar, este sistema evaluaba el riesgo de reincidencia para grupos, no para individuos; en tercer lugar, se basaba en datos a nivel nacional en lugar de datos locales de Wisconsin; en cuarto lugar, existían inquietudes sobre el potencial sesgo de los algoritmos de imposición de penas hacia minorías étnicas; y, en quinto lugar, el COMPAS fue diseñado para asistir al Departamento de Instituciones Penitenciarias, no necesariamente para su aplicación en el ámbito de los tribunales penales.

La decisión de utilizar software de IA en el proceso judicial ha generado escepticismo en la doctrina. Se critica que esta decisión no considera la presión sobre los jueces para usar estas herramientas ni los sesgos cognitivos que pueden surgir. Por ejemplo, los jueces tienden a

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

confiar en datos empíricos como estadísticas pese a que no sean completamente fiables. En el caso antes referido, Tribunal Supremo de Wisconsin fue criticado por no discernir adecuadamente al considerar que un acusado podía verificar la exactitud de un informe de evaluación de riesgos realizado por el software COMPAS. Es cierto que los datos eran públicos, pero el acusado no pudo verificar las variables influyentes en el resultado, ya que la empresa desarrolladora protegía sus secretos comerciales y no revelaba el funcionamiento del sistema. Por eso, el asunto *State v. Loomis* es un claro exponente de cómo los mecanismos de justicia algorítmica operan bajo un velo de opacidad, lo que impide que los litigantes cuestionen los resultados que producen. Es vital, por ello, supervisar estas herramientas, bajo el principio de neutralidad tecnológica, pues la IA de ayuda a la toma de decisiones corre el riesgo de generar discriminación o incluso crear nuevas formas de discriminación.

Esta falta de transparencia en el uso de *software* de IA en procesos judiciales plantea serias preocupaciones, ya que el acusado no puede defenderse adecuadamente si desconoce cómo se ponderaron las variables en su contra, ignorándose el sesgo de anclaje, que influye en la toma de decisiones judiciales. La falta de comprensión de los jueces sobre el funcionamiento del software aumenta el riesgo de seguir ciegamente sus resultados. Para abordar esto, es necesario mejorar la formación de los profesionales del derecho en tecnologías emergentes y aumentar la cooperación con investigadores y desarrolladores de IA. Estos desafíos subrayan la necesidad de mayor transparencia y conciencia sobre los posibles sesgos en el uso de IA en decisiones judiciales.

7. **Hacia una justicia algorítmica**

7.1. *Metamorfosis del ámbito jurídico*

La irrupción de las nuevas tecnologías está marcando un cambio significativo en el ámbito jurídico. Por parte del Poder Judicial la creciente influencia de la IA en el sistema judicial plantea interrogantes sobre su conformidad con el art. 6 CEDH, que garantiza el acceso a un tribunal y a un juicio imparcial e independiente. Si una hipotética "máquina de subsunción" pudiera tomar decisiones automatizadas basadas en el Derecho aplicable y los hechos del caso, se cuestionaría su compatibilidad con el CEDH. El acceso al tribunal estaría garantizado, pero la ausencia de un elemento humano en el proceso

de toma de decisiones no deja de plantear desafíos. Las máquinas carecen del sentido de justicia inherente a los humanos, lo cual no excluye que la IA pueda desarrollar habilidades emocionales en el futuro. La aplicación del Derecho va más allá de la simple interpretación de la ley escrita, requiriendo un análisis complejo de intereses jurídicos, doctrinas y jurisprudencia relevante. Por lo tanto, la IA podría ser adecuada solo para decisiones judiciales donde no haya discrecionalidad en las consecuencias jurídicas. Sería posible un proceso en el que un humano utilice la IA como apoyo, pero manteniendo la autonomía para tomar decisiones y anunciarlas requiriéndose para ello una definición clara de las funciones de la IA y un análisis detallado de la influencia del humano en la decisión final para garantizar el cumplimiento de los principios judiciales fundamentales.

La irrupción de las nuevas tecnologías está transformando la naturaleza del trabajo de los abogados, presentando desafíos y oportunidades. Un cambio notable es la influencia de las aplicaciones informáticas en tareas que tradicionalmente requerían juicio humano experto, como la predicción de resultados judiciales. Esto puede aumentar la transparencia jurídica y mejorar la eficiencia en la resolución de litigios y el acceso a la justicia. Sin embargo, también desafía el modelo tradicional de las empresas de Derecho privado, que se basan en horas facturables y el modelo socio-asociado. Las nuevas tecnologías ofrecen ventajas, permitiendo un trabajo más eficiente y una mayor especialización, proporcionando un valor agregado a los clientes.

En última instancia, estos avances tecnológicos están destinados a transformar tanto la forma en que los abogados llevan a cabo su trabajo como la manera en que abordan y resuelven los litigios en nombre de sus clientes. La capacidad de adaptación y la disposición a abrazar estos cambios serán fundamentales para el éxito en el panorama legal del futuro por eso los bufetes de abogados la utilizan para optimizar y simplificar procesos legales. Tecnologías como los contratos inteligentes, que monitorean la ejecución de los contratos y responden automáticamente a incumplimientos, están siendo cada vez más comunes, pero que dejan en el aire la pregunta sobre qué pasaría si una máquina representara a una parte ante un tribunal en lugar de un abogado humano. En el contexto de procesos penales, el derecho a la defensa está garantizado por el art. 6, ap. 3, letra c) CEDH. La representación por una máquina violaría el principio de un juicio justo, ya que se impediría sistemáticamente el acceso a un abogado defensor y aunque se pudiera considerar el uso de defensa técnica generada automáticamente por IA, esto no sería compatible con el principio de juicio oral, congénito a un juicio justo según el CEDH.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

En contraste, en procedimientos civiles, donde el litigio es iniciado por el demandante, el uso de IA se contempla desde una perspectiva diferente. Aquí, la IA podría generar alegatos escritos para simplificar los procedimientos, siempre y cuando se respeten los requisitos de oralidad. En procedimientos sin representación obligatoria por abogado, y si se cumplen los requisitos para renunciar a la oralidad, el uso de IA podría ser viable.

Ciertamente, la IA no puede reemplazar por completo a un abogado defensor en procesos penales, pero puede ser útil para apoyar actividades legales y en procedimientos civiles, la IA utilizarse en la fase preparatoria y en la generación de escritos, siempre que se respeten los requisitos legales y procesales pertinentes.

7.2. *Desafíos y preocupaciones*

La utilización de la IA en el ámbito legal plantea desafíos y preocupaciones que demandan un alto nivel de precisión y transparencia, así como la resolución de posibles sesgos o discriminaciones congénitas a los algoritmos. En este contexto, los profesionales del Derecho deben asumir un papel activo y crítico en el uso de esta tecnología, empleando su juicio y experiencia para complementar y contextualizar la información proporcionada por los sistemas de IA. Si se emplean de manera adecuada y ética, las herramientas de IA tienen el potencial de mejorar significativamente la eficacia y la calidad de la administración de justicia proporcionando a los abogados y otros operadores jurídicos recursos poderosos que les permitan desempeñar sus funciones de manera más eficiente y precisa. No obstante, resulta fundamental abordar los desafíos y preocupaciones asociados con el uso de la IA garantizando que estas tecnologías se implementen de manera que beneficien a la sociedad en su conjunto, evitando cualquier forma de perjuicio o inequidad. La responsabilidad de los profesionales del Derecho en este proceso tendrá la virtualidad de asegurar que la incorporación de la IA en el ámbito jurídico se realice de manera responsable, ética y orientada al bien común.

La implementación de la IA en la sociedad plantea desafíos que deben abordarse protegiendo los derechos humanos y las libertades fundamentales, evitando que pueda restringir la libertad ideológica, promoviendo un marco regulatorio que garantice los derechos de las personas y fomentando la innovación y la ética en el desarrollo algorítmico. En el ámbito educativo, deben implementarse programas

específicos sobre la IA y sus desafíos éticos, y la formación en Derecho debe incluir asignaturas dedicadas a la IA. La ética debe ser central en todo el proceso de desarrollo algorítmico, considerando aspectos como la transparencia, la "explicabilidad", los sesgos y la privacidad. Es crucial reflexionar sobre el propósito y las posibles consecuencias de la IA antes de su implementación. Las políticas y regulaciones deben diseñarse para prevenir el abuso de la IA y asegurar que su desarrollo se alinee con valores de justicia y equidad. Los desarrolladores y usuarios de IA deben estar comprometidos con la creación de sistemas justos y transparentes, evitando perpetuar desigualdades preexistentes, e implicando una reflexión constante sobre las implicaciones éticas y sociales de la IA.

La garantía de un uso ético y responsable de la IA amerita mecanismos de supervisión y rendición de cuentas que involucren a diversos actores, incluidos gobiernos, organizaciones de la sociedad civil y el sector privado capaces de identificar y corregir posibles desviaciones éticas en el uso de la IA, porque la integración de la ética en el desarrollo y uso de la IA no es solo una cuestión de responsabilidad técnica, sino un imperativo moral. Debemos asegurar que esta poderosa tecnología se utilice de manera que respete y promueva los derechos humanos, la libertad y la dignidad, construyendo así un futuro en el que la IA sea una fuerza para el bien común.

Es preciso que los abogados y las partes comprendan las limitaciones de la IA y no dependan exclusivamente de ella en decisiones legales, ya que la adopción de algoritmos predictivos plantea importantes cuestiones éticas y de transparencia sobre su funcionamiento y los datos utilizados para entrenarlos. Por esa razón la IA debe alinearse con el interés general, ser confiable, no aumentar desigualdades sociales y evitar que su evolución sea dictada exclusivamente por el mercado, lo que podría llevar a resultados inequitativos, con el concurso principal de cuatro acciones:

a) Abastecer a la IA de capacidad explicativa a través de una combinación de sistemas de aprendizaje profundo con IA simbólica, permitiendo mejorar la comprensión y la semántica en el procesamiento del lenguaje y la interpretación visual. La capacidad explicativa es fundamental para asegurar que las decisiones y recomendaciones de los sistemas de IA sean comprensibles y transparentes para los usuarios, lo que a su vez aumenta la confianza en estas tecnologías.
b) Proveer una IA más general y versátil, capaz de realizar múltiples tareas, lo cual requiere dotar a las máquinas de conocimientos

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

de sentido común, un problema complejo y de larga data en la IA. Los sistemas de IA actuales suelen ser altamente especializados y carecen de la flexibilidad y adaptabilidad que caracterizan a la inteligencia humana. Desarrollar una IA con habilidades de sentido común implicaría avances sustanciales en la representación del conocimiento y en la capacidad de razonamiento de las máquinas.

c) Abordar los aspectos éticos y sociales del desarrollo de la IA. Esto incluye garantizar que los sistemas de IA no perpetúen ni exacerben las desigualdades existentes, así como asegurar que su uso respete los derechos humanos y promueva la equidad y la justicia. Los desarrolladores de IA deben ser conscientes de los posibles sesgos en sus algoritmos y trabajar activamente para mitigarlos.
d) Proporcionar marcos legales y normativos que garanticen la gobernanza responsable de la IA, protegiendo a los individuos y a la sociedad en su conjunto de posibles abusos y mal uso de estas tecnologías.

Aún no se ha evaluado convenientemente la amenaza que representa la IA para las políticas públicas de los Estados y las instituciones democráticas, pero existe una preocupación legítima sobre la seguridad de los datos, la privacidad y el potencial pirateo de información sensible, especialmente en casos que involucran personalidades públicas y estatales. Siendo preocupaciones legítimas sobre la implementación de la IA en el ámbito jurídico, no pueden negarse sus potenciales beneficios, debiéndose abordar de manera adecuada los desafíos y riesgos asociados con su uso, mientras se aprovechan sus ventajas para mejorar la eficiencia, accesibilidad y equidad del sistema judicial.

7.3. *Necesidad de un marco regulatorio*

El acceso a la justicia constituye un pilar fundamental dentro del sistema jurídico, ya que su ausencia mina la integridad del debido proceso. En una sociedad marcada por un constante avance tecnológico, emergen herramientas como la IA que prometen facilitar las tareas humanas en diversos campos; sin embargo, para los Poderes Judiciales, la integración de esta tecnología aún representa un desafío significativo. Su ascenso se fundamenta en la convergencia de una mayor capacidad de procesamiento, la ampliación del acervo de datos

y los avances en algoritmos, consolidando así su posición como un concepto amplio que abarca algunos de los desarrollos tecnológicos más revolucionarios de nuestra era. La relevancia de la IA se desprende tanto de las oportunidades que brinda como de los desafíos que plantea.

Sin descuidar las aplicaciones de la IA tendentes a impulsar el crecimiento económico y a mejorar la eficiencia, no debe olvidarse que también suscitan importantes riesgos e incertidumbres. De ahí la importancia de implementar un sistema efectivo de justicia algorítmica que identifique y erradique los sesgos computacionales discriminatorios de manera rápida y eficiente. Este sistema: a) promovería un acceso más amplio a la justicia; b) aumentaría la confianza en la tecnología de IA; c) establecería medidas para medir y comunicar el progreso hacia la eliminación de los sesgos indeseables; d) proporcionaría un marco para el desarrollo de políticas y regulaciones significativas para la IA; y e) facilitaría el cumplimiento y la litigación contra la discriminación algorítmica ilegal. Como puede observarse, se trata de una compleja tarea cuyo resultado sería nuevo modelo hacia un sistema de justicia algorítmica.

La justicia algorítmica puede mejorar la accesibilidad a la justicia y acelerar la toma de decisiones, pero amenazar paralelamente otras garantías consustanciales al derecho a un juicio justo, al poner en riesgo el deber de motivación, la independencia e imparcialidad de los jueces y la igualdad de armas entre las partes (Nieva 2018, 128–130). De ahí la necesidad de una gobernanza adecuada de la IA para guiar el desarrollo, despliegue y mantenimiento de las tecnologías que la conforman de manera ética y responsable. Incluye principios, normas y marcos que abordan aspectos como la ética, la transparencia, la responsabilidad y la gestión de riesgos y su finalidad principal es asegurar el uso ético de la IA mitigando los riesgos asociados, como la parcialidad y las violaciones de la privacidad. Una gobernanza sólida generará confianza entre los usuarios y las partes interesadas al garantizar que las tecnologías de IA se utilicen de manera beneficiosa y cumplan con las expectativas legales y sociales. Pero requiere, en primer lugar, el establecimiento de marcos regulatorios que equilibren la innovación con la protección del interés público y los derechos individuales; en segundo lugar, la revisión y actualización de las normativas relacionadas con la responsabilidad civil, los derechos de los consumidores y la protección de datos; y, por último, la implementación de modelos de IA explicables y transparentes, abordando los desafíos éticos y sociales asociados con esta tecnología para que los usuarios puedan comprender y confiar en sus decisiones

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

(Martín Diz 2021, 65–85; Barona 2021). Entre estos requerimientos ocupa un lugar prioritario la consideración de que la perspectiva de género en la IA y la concienciación social sobre sus implicaciones son etapas a superar para garantizar un uso responsable y ético de esta tecnología pues, de lo contrario, podría tener consecuencias negativas para la humanidad.

La garantía del uso ético y seguro de la IA parte de una regulación adecuada para asegurar su uso ético y seguro y del establecimiento de marcos normativos que definan los límites y responsabilidades de los desarrolladores, fabricantes y usuarios de la IA, siendo fundamentales la transparencia y la responsabilidad para generar confianza en la IA, por lo que se deben establecer mecanismos para que los desarrolladores y proveedores de servicios sean responsables de las decisiones tomadas por sus sistemas a través de evaluaciones éticas para garantizar que la IA se utilice de manera no discriminatoria. En su consecución, resulta conveniente, de un lado, la creación de comités de ética y de mecanismos de revisión que protejan los derechos humanos y los valores éticos y, de otro lado, la implementación actual de la justicia algorítmica para que asuma las garantías de equidad necesarias según el principio de no discriminación y el derecho a un proceso equitativo, tal como se define en la jurisprudencia del TEDH, del TJUE y del Tribunal Constitucional.

En suma, es fundamental implementar un sistema efectivo de justicia algorítmica con el objeto de:

a) Identificar y erradicar los sesgos computacionales discriminatorios de manera rápida y eficiente. Con ello no solo se promoverá un acceso más amplio a la justicia y aumentaría la confianza en la tecnología de IA, sino que también se establecerán medidas para medir y comunicar el progreso hacia la eliminación de dicho sesgo, proporcionado un marco para el desarrollo de políticas y regulaciones significativas para la IA, facilitando el cumplimiento y la litigación contra la discriminación algorítmica ilegal.
b) Asegurar que la integración de la IA en el sistema judicial cumpla con los estándares legales de justicia y equidad. El desafío futuro consiste en maximizar los beneficios de la IA para la sociedad al tiempo que se mitigan sus riesgos, fomentando la innovación y equilibrando los intereses sociales y ello plantea la necesidad de definir qué valores guiarán el desarrollo tecnológico y qué principios del Estado de Derecho brindarán una base sólida para este propósito como una prioridad para preservar los valores fundamentales de la sociedad.

Bibliografía

Añón, María José. 2018. «El derecho de acceso como garantía de justicia: perspectivas y alcance», *Acceso a la justicia y garantía de los derechos en tiempos de crisis: de los procedimientos tradicionales a los mecanismos alternativos*, coordinado por Cristina García–Pascual, 19-75. Valencia: Tirant lo Blanch.

Ashley, Kevin D. 2017. *Artificial Intelligence and Legal Analytics. New Tools for Law Practice in the Digital Age*, Cambridge: Cambridge University Press.

Barona, Silvia. 2021. *Justicia algorítmica y neuroderecho. Una mirada multidisciplinar,* Valencia: Tirant lo Blanch.

Buiten, Miriam C. n.d. «Towards intelligent regulation of artificial intelligence.» *European Journal of Risk Regulatio*n 10 (1): 41–59.

Carrizo, Adán. 2019. «El acceso a la justicia de las personas en condición de vulnerabilidad: un reto pendiente para los derechos humanos», *Los Derechos Humanos 70 años después de la Declaración Universal*, dirigido por Nieves Sanz, 287-310. Valencia, Tirant lo Blanch.

Castillejo, Raquel. 2022. «Digitalización y/o inteligencia artificial», *Inteligencia artificial legal y administración de justicia,* dirigido por Sonia Calaza y Mercedes Llorente, 55-90. Cizur Menor: Thomson–Reuters–Aranzadi.

Chatterjee, Payel, Aman Singhania y Yuvraj S. Sharma. 2023. *Technology and artificial intelligence: Reengineering arbitration in the new world*, International Bar Association. Arbitration Commitee Articles, 20 de diciembre. Acceso el 17 de septiembre de 2024 https://www.ibanet.org/tchnology-and-artificial-intelligence-reengineering-arbitration-in-the-new-world.

Ebers, Martin y Paloma Krõõt. 2023. *Artificial intelligence and machine learning powered public service delivery in Estonia: opportunities and legal challenges,* Cham: Springer.

Esparza, Iñaki. 2022. «Derecho fundamental a la protección de datos de carácter personal en el ámbito jurisdiccional e inteligencia artificial. en especial la LO 7/2021, de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales», *Inteligencia artificial legal y administración de justicia,* dirigido por Sonia Calaza y Mercedes Llorente (181–209). Cizur Menor: Thomson–Reuters–Aranzadi.

Fernández Pérez, Ana. 2024. «Tránsito del arbitraje de las tecnologías de la información a la inteligencia artificial en las controversias internacionales», *Anuario de Arbitraje 2024*, editado por Mª José Menéndez, Cizur Menor: Aranzadi.

Fernández Rozas, José Carlos. 2024. «La ley de inteligencia artificial de la Unión Europea: un modelo para innovaciones radicales, responsables y transparentes basadas en el riesgo», *La Ley: Unión Europea* 124, abril.

Forrest, Katherine B. 2021. *When machines can be judge, jury, and executioner: justice in the age of artificial intelligence*, New Jersey: World Scientific.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 203-235
https://doi.org/10.18543/djhr.3195 • http://djhr.revistas.deusto.es/

Gómez Colomer, Juan L. 2022. «Derechos fundamentales, proceso e inteligencia artificial: una reflexión», *Inteligencia artificial legal y administración de justicia,* dirigido por Sonia Calaza y Mercedes Llorente, 257-287. Cizur Menor: Thomson–Reuters–Aranzadi.

Greenstein, Stanley. 2022. «Preserving the rule of law in the era of artificial intelligence (AI)», *Artificial Intelligence and Law* 30: 291–323.

Leslie, David, Christopher Burr, Mhairi Aitken, Josh Cowls, Mike Katell y Morgan Briggs. 2021. *Artificial intelligence, human rights, democracy, and the rule of law: a primer*. Council of Europe, The Alain Turing Institute.

Marrow, Paul B., Mansi Karol y Steven Kuyan. 2020. «Artificial intelligence and arbitration: the computer as an arbitrator - are we there yet?», *Dispute Resolution Journal* 74 (4): 35–76.

Martín Diz, Fernando. 2021. «Modelos de aplicación de Inteligencia Artificial en justicia: asistencial o predictiva versus decisoria», *Justicia algorítmica y Neuroderecho. Una mirada multidisciplinar,* dirigido por Sonia Calaza y Mercedes Llorente, 65-85. Cizur Menor: Thomson–Reuters–Aranzadi.

Mercan, Gamze y Zumrut V. Seiçuk. 2024. «Artificial intelligence (AI) activities in legal practices», *International Journal of Eurasian Education and Culture* 9 (25): 131–144

Nieva, Jordi. 2018. *Inteligencia artificial y proceso judicial*, Madrid: Marcial Pons.

Ortiz de Zárate, Lucía y Ariana Guevara. 2021. «Inteligencia artificial e igualdad de género. Un análisis comparado entre la UE, Suecia y España», *Estudios de Progreso* 101, Fundación Alternativas. Acceso el 17 septiembre 2024, https://www.igualdadenlaempresa.es/recursos/estudiosMonografia/docs/Estudio_Inteligencia_artificial_e_igualdad_de_genero_Fundacion_Alternativas.pdf

Re, Richard M. y Solow–Niedermann, Alicia. 2019. «Developing artificially intelligent justice», *Stanford Technology Law Review* 22 (2): 242–289.

Sartor, Giovanni. 2020. «Artificial intelligence and human rights: Between law and ethics». *Maastricht Journal of European and Comparative Law* 27 (6): 705–719.

Schäferling, Stefan. 2023. *Governmental Automated Decision–Making and Human Rights: Reconciling Law and Intelligent Systems*. Cham: Springer.

Scherer, Maxi. 2019. «Artificial intelligence and legal decision–making: The wide open?», *Journal of International Arbitration* 36 (5): 539–573.

Završnik, Aleš, 2020. «Criminal justice, artificial intelligence systems, and human rights», *ERA Forum: Journal of the Academy of European Law* 20 (4): 567–583.

Ética en crisis: El impacto de la carrera armamentística de las armas autónomas en nuestros valores morales

Ethics in crisis: The impact of the autonomous arms race on our moral values

Jorge Couceiro Monteagudo
Universidad Complutense de Madrid. España
jorgecouceiromonteagudo@gmail.com
ORCiD: https://orcid.org/0009-0005-0608-2304

https://doi.org/10.18543/djhr.3083

Fecha de recepción: 31.05.2024
Fecha de aceptación: 25.09.2024
Fecha de publicación en línea: diciembre de 2024

Cómo citar / Citation: Couceiro, Jorge. 2024. «Ética en crisis: el impacto de la carrera armamentística de las armas autónomas en nuestros valores morales». *Deusto Journal of Human Rights*, n. 14: 237-257. https://doi.org/10.18543/djhr.3083

Sumario: Introducción. 1. ¿Qué ha cambiado con la autonomía? 2. ¿Nos deshumaniza? Tecnificación de las operaciones humanas. 3. ¿Cuáles son las implicaciones en nuestros valores humanos? 4. ¿Hay posibilidad de retroceder? Conclusión. Referencias.

Resumen: Las matanzas sistemáticas han sido parte de los episodios más oscuros de la historia. Actualmente, se presentan como un objetivo en la guerra, especialmente con la IA militar y las armas autónomas. Sus defensores afirman que estas armas superarán a los humanos en capacidad militar y moral, ofreciendo violencia más precisa y sin emociones. Sin embargo, argumentamos que estas armas replican e intensifican los desafíos morales del pasado. La violencia autónoma desvaloriza moralmente a las víctimas y reduce la responsabilidad moral de los atacantes, poniendo en peligro las restricciones al uso de la fuerza militar.

Palabras clave: Armas autónomas letales, inteligencia artificial, valores morales, impacto, ética, crisis, violencia.

Abstract: Systematic killings have been part of the darkest episodes in history. Today, they are presented as a goal in warfare, especially with military AI and autonomous weapons. Proponents claim that these weapons will surpass humans in military capability and morale, offering more precise and

emotionless violence. However, we argue that these weapons replicate and intensify the moral challenges of the past. Autonomous violence morally devalues victims and reduces the moral responsibility of attackers, jeopardising restrictions on the use of military force.

Keywords: Lethal autonomous weapons, artificial intelligence, moral values, impact, ethics, crisis, violence.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 237-257
https://doi.org/10.18543/djhr.3083 • http://djhr.revistas.deusto.es/

Introducción

Gracias al avance de la tecnología, a día de hoy es posible que, sin saberlo e indirectamente, podamos formar parte, como engranajes en una máquina, de acciones cuyos efectos están lejos de lo que nosotros podemos percibir como acciones propias. Fuera de nuestros ojos e imaginación, si pudiéramos concebir sus consecuencias, no podríamos aprobarlas. Este hecho ha cambiado los fundamentos mismos de nuestra existencia moral (Anders 1962). En este instante nos enfrentamos a un futuro acelerado, en el que la creciente tendencia hacia la tecnificación humana (la sustitución del trabajo humano por la tecnología) y la sistematización exacerbaría la aplicación deshumanizada de la fuerza letal y conduciría a más violencia. En la actualidad, nos enfrentamos a un futuro acelerado con sistemas de armas autónomas dentro de una guerra altamente sistematizada. En particular, la velocidad del despliegue y la escala a la que pueden llegar los sistemas de armas habilitadas por IA deberían provocar una reflexión sobre las implicaciones morales sobre la integración de lógicas y sistemas no humanos en los procesos existentes de violencia militar.

El uso de sistemas autónomos letales (LAWS) en el acto de matar constituye siempre una forma sistematizada de violencia. En este proceso, toda la cadena de mando —desde el comandante hasta el operador y el objetivo— se ve sometida a una tecnificación que no solo degrada los valores morales hacia las víctimas, sino que también erosiona la agencia moral de quienes participan en la implementación de esta violencia autónoma. Esta dinámica amenaza con debilitar las restricciones fundamentales que legitiman el uso de la fuerza militar.

En este contexto, la reciente proliferación de estudios que defienden los LAWS como una alternativa moralmente superior —incluso más humana o posthumanista— para la administración de la fuerza letal, plantea un desafío significativo. Estas ideas, no obstante, permiten cuestionar los beneficios aparentes de esta tecnología, que suelen fundamentarse en una visión abstracta y excesivamente idealizada de su funcionamiento en los entornos dinámicos y complejos de la guerra.

1. ¿Qué ha cambiado con la autonomía?

La organización militar y la lucha bélica han sido ordenadas y reordenadas a lo largo de la historia en sistemas más fijos e instrumentales (Keegan 1994). Las reglas de la guerra también se han

normalizado para vincular a los combatientes a un conjunto más fijo de medidas que limitan el alcance de la violencia permisible. La autonomía, o la sistematización intensificada, ha cambiado los modos de violencia en los que la lógica del cálculo, la clasificación y la optimización del acto de eliminación se vuelven primordiales. Este modo de ver la violencia pone en peligro las restricciones morales esenciales sobre el uso de la fuerza y es intrínseca a las armas autónomas letales habilitadas por la IA. De este modo, las LAWS reproducen, y en algunos aspectos intensifican, los desafíos morales asociados con episodios anteriores de asesinatos sistemáticos intensificados.

Las nuevas tecnologías pueden alterar el equilibrio de la guerra de distintas maneras. En algunos casos, la ventaja tecnológica disruptiva no resuelve los problemas de la guerra, sino que los hace más prominentes y duraderos (Crootof 2019). El desafío moral de la violencia autónoma es un claro ejemplo de esto. La matanza sistemática en la guerra no es nueva ni exclusiva de esta tecnología; sin embargo, los sistemas autónomos aceleran muchas de sus peores características debido a sus particularidades. Y, si bien su mayor desafío no radica únicamente en la falta de humanidad, esta tecnología plantea serias implicaciones éticas y operativas. Los seres humanos seguirán siendo intrínsecos a estos sistemas; lo que está en juego es el tipo de humanidad que esta tecnología hace cada vez más probable. Las armas autónomas nos liberan de una conducta llena de pasión y volatilidad propia del ser humano, corriendo el riesgo de hundirnos aún más en una conducta fría y desapasionada de los sistemas humanos.

A medida que la guerra se ha vuelto más compleja, los identificadores tradicionales ya no hacen que el enemigo sea coherentemente legible y visible, y cada vez más los datos sirven como sustitutos. "En condiciones tan difíciles de ilegibilidad y difluencia […] los datos se buscan en un grado sin precedentes" para identificar y rastrear enemigos y predecir quien podría convertirse en uno (Ansorge 2016, 124). Este método se amplifica con LAWS, donde los sistemas de IA comprenden e identifican objetivos basándose únicamente en el reconocimiento y la clasificación de objetos a través de redes neuronales. La IA representa el mundo tal como lo percibe, como un conjunto de datos y patrones relacionados a partir de los cuales se pueden predecir y calcular los resultados, incluida la decisión sobre cuáles deben de ser los objetivos. Y es que la razón por la cual un individuo puede ser asignado como objetivo para ser eliminado llega a conocerse mediante la probabilidad estadística en el que los fenómenos discretos e inconexos se unen y se evalúan

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 237-257
https://doi.org/10.18543/djhr.3083 • http://djhr.revistas.deusto.es/

correlativamente (Cheney-Lippold 2019), dejando de lado quién son, cómo se comportan o qué pretenden. Dentro del proceso de selección, los datos se reconfiguran continuamente para ajustarse a modos específicos de clasificación. A partir de estos datos, el sistema calcula una inferencia sistemática de quién, o qué, se identifica con un patrón de normalidad o anormalidad para eliminar la amenaza.

Este nuevo modo de identificar al enemigo está plagado del riesgo de ver patrones y hacer inferencias donde no las hay, un desafío bien conocido en el razonamiento humano que se organiza en estructuras algorítmicas y se sistematiza. Un sistema de IA encargado del reconocimiento de imágenes procesa una imagen como un conjunto de píxeles, y cada píxel como una serie de valores que representan distintas propiedades. En otras palabras, una 'matriz de números que corresponden al brillo y el color de los píxeles de la imagen' (Mitchell 2019, 78). Con el fin de entrenar tales sistemas para identificar a un enemigo, el sistema primero tendría que ser entrenado con un número considerable de mágenes debidamente etiquetadas (por ejemplo, con las variables "enemigo" o "terrorista") introducidas en sus parámetros. Mediante las redes convencionales, ciertas características de la imagen se establecen para clasificar el objetivo en el que se entrena como útiles (Anders 1962). Estos datos se introducen en una red neuronal, que ordena y clasifica la entrada para predecir qué objeto representa la imagen con un cierto grado de confianza, expresado en valores porcentuales. Aunque esto nunca es un reflejo verdadero y completo de la realidad.

"Es sistemáticamente difícil para las LAWS clasificar un evento u objeto en una categoría particular. En cambio, su proceso lo revisarán y diseccionarán de acuerdo con un número inapropiadamente pequeño de características" (Walker 2021, 17 y 14). Para que un arma sea eficaz en el contexto dinámico de la guerra, esta debe "calcular continuamente nuevas probabilidades para su escenario inmediato" (Anders 1962, 1), un proceso que está gestionado por una función de error. Es decir, el método para tomar una decisión de eliminación con LAWS es aquella que se basa en la aproximación, la racionalización y el suavizado de los puntos de datos. Dentro de este proceso, las personas se convierten no solo en objetos en la aplicación selectiva de la violencia, sino en objetos que se constituyen a través de patrones algorítmicos. Primero se identifican patrones y se trazan líneas de asociación y, en base a estos, se realizan cálculos de muerte. La lógica intrínseca de la IA se basa en esta clasificación y codificación de la vida en datos computables para identificar objetos y patrones entre objetos. "Ser inteligible para un modelo estadístico que es transcodificado en un

marco de objetivación y llega a ser definido, calculado de forma cruzada, como un objeto computacionalmente determinado y procesable (Cheney-Lippold 2019, 513-535). Esta fundamentación asistemática produce no solo una pura objetivación, sino también, si el objetivo es humano, una subjetivación y desindividualización. Estos individuos "no pueden confiar en nada único para ellos porque la solidez de su subjetividad está determinada totalmente fuera de uno mismo, y de acuerdo con lo que sea que se incluya dentro de una clase de referencia o conjunto de datos no pueden decidir" (Anders 1962, 1).

La tecnología de las armas autónomas ha avanzado de forma tan abrupta en los últimos años que se prevé que continúe haciéndolo en los próximos años, ya que está planteando una ventaja estratégica tan importante que muchos estados no pueden omitirla por el potencial militar que supone. Un potencial derivado de las sofisticadas innovaciones de IA a través de redes neuronales y de aprendizaje automático, junto con mejoras en la potencia de procesamiento de las computadoras, han abierto un campo de posibilidades en la amplia gama de aplicaciones militares, como la toma de decisiones autónomas, incluida la selección de objetivos. El aspecto que más destaca de esta tecnología es el potencial del sistema de armas para seleccionar y atacar objetivos de forma autónoma, sin intervención o acción humana. La definición ofrecida por el Comité Internacional de la Cruz Roja dice así: "Cualquier sistema de armas con autonomía en sus funciones críticas. Es decir, un sistema de armas que pueden seleccionar (buscar, identificar, rastrear o seleccionar) y atacar (usar la fuerza, contra, neutralizar, dañar o destruir) objetivos sin intervención humana" (Heyns 2017, 46–71; Asaro 2012, 687–709). A diferencia de los drones que son operados a distancia, las armas autónomas relegan al ser humano como supervisor en el bucle de la toma de decisiones o cadena de mando *(humans on the loop)*, o eliminan al humano por completo *(humans out of the loop)*. Cuando el humano permanece fuera de la toma de decisiones, las decisiones y acciones podrían iniciarse y completarse de forma autónoma, basándose en datos de entrada, sensores, algoritmos y programas de software.

Tanto en la teoría como en la práctica, las LAWS pueden acortar el tiempo de reacción entre el sensor y el tirador de minutos a segundos. La capacidad de procesar grandes cantidades de datos complejos en un marco de tiempo acelerado se considera un beneficio estratégico significativo, incluso si se produce sin la supervisión humana directa. Esto nos lleva a que la toma de decisiones de un agente humano pase de ser decisiones individuales para elegir dentro de la capacidad humana qué grupo de robots deben participar para después tomar ellos

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 237-257
https://doi.org/10.18543/djhr.3083 • http://djhr.revistas.deusto.es/

las decisiones individuales (Knight 2021). Esta nueva forma de ejecutar nos proporciona una visión más amplia de una guerra futura habilitada por IA en su totalidad. Una IA interconectada, que cruce dominios y que comprima el tiempo. Este nuevo concepto llamado "red de redes" puede llegar a convertirse en el nuevo modelo de los conflictos futuros, que puede requerir que se tomen decisiones en horas, minutos o potencialmente segundos en comparación con el proceso de varios días para analizar el entorno operativo y emitir órdenes. Es una visión que sistematiza todos los procesos y operaciones, incluidos los objetivos letales de la guerra, en la que tanto la velocidad como la escala prevista priorizan claramente la violencia autónoma. Aquí los LAWS basados en IA serán el punto clave para hacer realidad estas visiones de futuro (Congressional Research Service 2022).

Actualmente, hay varios tipos de LAWS, que incluyen municiones merodeadoras habilitadas por IA con armas equipadas con sistemas de enjambre de drones armados con la capacidad de identificar amenazas basadas en ciertos parámetros de entrada y salida, con la finalidad de fijar ciertos objetivos y eliminarlos una vez que se ha alcanzado un cierto valor condicionante. Sin embargo, este tipo de armas necesitaría ser entrenado mediante grandes cantidades de datos que sean relevantes para una zona de conflicto o área de participación y requiere actualizaciones frecuentes, ya que el campo de batalla es cambiante y da lugar a nuevos parámetros o datos ligeramente heterogéneos a los datos bajos los cuales el arma ha sido entrenada, que puede confundir al sistema (Walker 2021).

2. **¿Nos deshumaniza? Tecnificación de las operaciones humanas**

Dentro de la tecnificación del ser, es necesario apuntar que no solo las víctimas son las que se ven afectadas negativamente por los asesinatos sistemáticos. Los que forman parte como participantes también pueden verse abrumados o reducidos por la lógica de los sistemas que gobiernan la organización y la imposición de la violencia. En un contexto de matanza sistemática, muchas de las características que desearíamos preservar y cultivar —el juicio, la responsabilidad personal, la autorreflexión, la moderación moral, etc.— tienen apenas espacio para operar. Lo que define este espacio es la lógica de la eficiencia y la velocidad en la que el ser humano, siendo operador, tiene la tarea de trabajar dentro de la lógica del sistema correspondiente. Este desafío se evidencia en la violencia autónoma, un método de asesinato que genera, al igual que sus antecedentes

históricos, problemas relacionados con la autorización, la rutinización y la deshumanización.

Uno de los puntos más importantes, que está relacionado con la autorización, son las estructuras de autoridad, siendo uno de los pilares que autorizan y sostienen los asesinatos en masa. Los sistemas computacionales exigen la deferencia de los operadores o comandantes, que rara vez comprenden completamente los procesos involucrados en la toma de decisiones computacionales. Ante tal complejidad y abstracción, a los seres humanos no les queda más remedio que confiar en la superioridad cognitiva y racional de esta autoridad clínica. Dicha relación se conoce como "sesgo de automatización" y es un fenómeno bien documento en la literatura sobre las interacciones hombre-máquina (Cummings 2004). De este modo, la autoridad tecnológica sirve para relajar las tensiones morales (Emery 2016). Volviendo a las LAWS, lo que está en juego no es un proceso formal y jerárquico de autorización, sino uno que emerge del carácter ostensiblemente neutral y superior de la propia máquina. Independientemente del rango, la capacidad de desafiar la autoridad de la lógica de la máquina se ve debilitada.

Dentro de un entorno digital complejo, la experiencia, la cognición y la acción humana se ven mediadas y moderadas a través de la lógica de las máquinas. Dentro de este marco, la posibilidad de ejercer un albedrío moral se ve determinantemente truncado tanto para los comandantes como para los operadores (Schwarz 2021). Dicho de otro modo, la capacidad de decidir se ve influida mediante el entorno de control distribuido, relevante para el control humano de las LAWS (Ekelhof 2019). Sin embargo, el efecto es particularmente relevante cuando se pide a los operadores que actúen sobre la información entrante, incluida la información de vida o muerte, en cuestión de segundos. Los operadores ante tal nivel de estrés por dichas limitaciones pueden carecer tanto de un "nivel suficiente de conciencia situacional para hacer juicios significativos" como de "suficiente información sobre los parámetros bajo los cuales las partes automatizadas o autónomas de los módulos de mando que seleccionan y priorizan las amenazas para examinar la selección de objetivos y abordar el ataque" (Bode et Watts 2021, 28). "Como operadores en el bucle, el humano se convierte así en el módulo .exe[1] en la red computacional más amplia, con una capacidad limitada, si

[1] En programación, un archivo .exe se conoce como archivo ejecutable. Un archivo de este tipo contiene instrucciones codificadas para un proceso computacional que se activa mediante un usuario u otro evento.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 237-257
https://doi.org/10.18543/djhr.3083 • http://djhr.revistas.deusto.es/

existe, para anular o intervenir en la acción preestablecida" (Corbett 2008, 167). En este contexto, es altamente probable que la combinación del despliegue de LAWS y la percibida superioridad de la lógica de la máquina genere una situación en la que el personal militar "se involucre en una acción sin considerar las implicaciones de esa acción y sin tomar realmente una decisión" (Kelman 1973, 25-61). A pesar de ello, esto no quiere decir que los agentes humanos sean una salvaguarda infalible contra las malas acciones, y tampoco liberas a las LAWS de sus aspectos moralmente problemáticos. Es necesario que los acusados de cometer actos de violencia comprendan el contexto y las consecuencias de sus acciones, y sean capaces de reconocer cuándo deben ceder ante la violencia y tengan la capacidad de actuar de acuerdo con este impulso, en lugar de retirarse del proceso. El peligro, por lo tanto, de los sistemas de Armas Autónomas Letales (SAAL) es que carecen de la capacidad misma para cumplir con estos estándares.

Otro aspecto de la tecnificación es la rutinización, que opera tanto a nivel individual como organizacional, cambiando el enfoque hacia lo puramente procedimental. A nivel individual, el operador es un elemento funcional dentro de la lógica del sistema, que ejecuta tareas específicas con una visión limitada de la situación. A nivel organizativo, las tareas relacionadas con la acción están divididas y, a menudo, difusas. El crear una rutina facilita la eficiencia, la precisión del procedimiento, la velocidad de reacción y ejecución de la tarea, etc. A su vez, se convierte en la norma, el estándar para llevar a cabo bien la acción, y la "naturaleza de la tarea se disocia por completo de la realización misma" (Anders 1962, 1). El principal peligro de la rutinización de la violencia es que excluirá las oportunidades de intervención moral y, por lo tanto, debilitará la restricción moral (Emery 2022).

En los conflictos, existen ambigüedades que siguen sin estar resueltas, en las que no se pueden establecer certezas en cuanto a la identidad y la responsabilidad de un objetivo potencial. Incluso cuando poseemos un conjunto de parámetros para tomar tales determinaciones con una confianza razonable, persiste cierto grado de incertidumbre. Y es esta incertidumbre la que deja espacio para el razonamiento ético, que después permite la intervención ética. Tales intervenciones son necesarias cuando el sistema y las reglas son estructuralmente inclusivas, lo que obliga a seleccionar a aquellos que han sido categorizados falsamente como objetivos legítimos. Las LAWS son idealizadas por algunos como sistemas potenciales capaces de "proezas éticas" (Umbrello et al. 2019, 273–282) diferenciando la parte cognitiva de la emotiva, dejándonos con agentes de violencia con menos poder moral.

A su vez, es importante mencionar que este desafío no se limita al extremo violento de la cadena de muerte. La rutina computacional incorporada en las LAWS reduce el espacio para la agencia humana. Los seres humanos permanecen dentro del sistema, pero la responsabilidad de la fuerza letal se difumina, o se separa, a través del propio proceso del sistema. En el peor de los casos, este desapego de responsabilidad facilita la aplicación descuidada o incluso deliberada de la violencia injusta.

Como último punto, respondiendo a la pregunta de la deshumanización, cuando se aplica un enfoque sistemático de la matanza, la deshumanización se multiplica. Las víctimas son cosificadas y despojadas de los derechos y el reconocimiento que de otro modo se les debería otorgar en virtud de su condición de seres humanos. Aunque es importante mencionar que este proceso también suele afectar deshumanizando al perpetrador. La deshumanización del soldado, del operador y de aquellos que establecen los parámetros para matar se afianza gradualmente a medida que funciona dentro del sistema más amplio de matar en el que la cognición y el afecto se separan. Cuando la responsabilidad personal, las relaciones humanas y la empatía se descartan sistemáticamente, uno no puede actuar como un ser humano moral (Kelman 1973). Como revela la rica literatura sobre este tema, estos procesos invitan a la violencia y al abuso (LeMoncheck 1985, Nussbaum 1995, Zurbriggen 2013).

¿Estos desafíos son inherentes a esta tecnología o son potencialmente resolubles? Según algunos, muchos de los problemas que asociamos con las LAWS pueden mitigarse si hacemos que los humanos sean funcionalmente relevantes dentro del sistema de toma de decisiones. Es decir, se inclinan hacia la tecnificación del ser. Los humanos deben ser capaces de entender y trabajar con la lógica de la máquina para obtener el mejor resultado: "Humano débil + máquina + mejor proceso [es] superior a [cualquier] computadora [o]... un humano fuerte + máquina + proceso inferior" (Kasparov, citado en Phillips-Levine et al. 2022, 1). Esto es intuitivo para la guerra futura solo si aceptamos que la lógica del sistema debe prevalecer en el proceso de matar con las LAWS. Un futuro así sancionaría a los sistemas de violencia que presentan a los enemigos ineludiblemente como objetos inhumanos, y hacen que los combatientes sean cada vez más inertes moralmente. Las leyes aceleran y racionalizan la decisión de matar, pero al hacerlo, abren nuevos espacios para infracción moral. La violencia computacional y desapasionada de las armas autónomas es incompatible con la regulación y la "humanización".

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 237-257
https://doi.org/10.18543/djhr.3083 • http://djhr.revistas.deusto.es/

3. **¿Cuáles son las implicaciones en nuestros valores humanos?**

Las matanzas sistemáticas en la historia son un gran debate sobre la moralidad de las armas autónomas. A menudo el debate se centra en la función y el valor de la humanidad. Es un debate disperso, pues la imagen que debería cumplir un soldado según el Derecho Internacional Humanitario es la de un combatiente justo, sin embargo, esto es un tipo ideal de soldado, que está en desacuerdo con gran parte de la experiencia humana en guerras pasadas y presentes (Williams 2021). Los combatientes humanos internalizan las reglas del campo de batalla con demasiada lentitud y las descartan demasiado rápido para que sean consistentemente efectivos. Durante el paso del tiempo en el que se ha librado la guerra, los participantes humanos, impulsados por la rabia, el miedo y el odio, han cedido a sus 'pasiones locas' y han aterrorizado y asesinado a partes inocentes (Best 1980). Es esta imagen de la humanidad la causa de la misera en la guerra a la que hacen referencias los defensores de las armas autónomas cuando enmarcan la tecnología como una alternativa éticamente superior (Riesen 2022).

Y aunque no esté equivocada, esta descripción pesimista de la humanidad es excesivamente estrecha, excluyendo otros tipos de mala conducta e inmoralidad impulsadas por el ser humano que es más probable que los sistemas de armas autónomas potencien, a que las prevengan. La atrocidad de la guerra a menudo tiene su origen en la pasión: el odio al enemigo y la euforia y la alegría de su sufrimiento. Sin embargo, además de esto, existen las crueldades más desapasionadas y sistemáticamente dispensadas. "La violencia fría debería perturbarnos mucho más que la bestia de la ira en el hombre" (Glover 2000, 64). Matar de modo tan frío está impulsado menos por la animadversión personal que por un cálculo lógico. Los modelos sistemáticos y desapasionados de 'control de plagas' de matanza han sido una característica de algunos de los episodios más destructivos de la historia humana (Münkler 2004).

El problema del asesinato sistemático se deriva principalmente de su relación histórica con el daño inhumano e injusto; es decir, el uso de sistemas organizados para llevar a cabo matanzas masivas de personas inocentes. Sin embargo, es importante destacar que muchos también se sienten repelidos por el proceso de asesinar sistemáticamente y por el grado en que la subsunción de las emociones humanas en sistemas intensificados de violencia socava el estatus moral tanto de ellos que dispensan violencia como de los que reciben el daño. Mostrar esta historia y las realidades empíricas de la violencia en la guerra ayuda a ir

más allá de los debates sobre las armas autónomas, excesivamente impregnados de supuestos teóricos abstractos.

Los procesos asociados con el asesinato sistemático, en las versiones más intensificadas, ponen en peligro la restricción al uso de la fuerza. Esto se observa en primer lugar en la relación con la situación de las personas a las que se dirige. La sistematización impone o incentiva categorías totalizadoras que suprimen las diferencias individuales de los destinatarios, incluidas aquellas que podrían ser relevantes para nuestro juicio moral sobre la legitimidad de los objetivos. Estos procesos han estado presentes en gran parte de la violencia colonial de los siglos anteriores. "El colonialismo estableció la idea de poblaciones enteras como objetivos legítimos" (Freedman 2017, 36). Bajo este contexto, se fijó la categorización, negando a las personas seleccionadas la oportunidad de expresar su inocencia y su inmunidad frente a daños directos y deliberados. Las tácticas británicas durante la Segunda Guerra ejemplifican este proceso. "Los británicos debían expulsar a los guerrilleros en una seria de campañas sistemáticas, organizadas como un tiro deportivo, con el éxito definido en una bolsa semanal de muertos, capturados y heridos, y barrer el país sin todo lo que pudiera dar sustento a los guerrilleros, incluidas las mujeres y los niños… Fue la eliminación de civiles, el desarraigo de toda una nación, lo que dominaría la última fase de la guerra" (Pakenham 1979, 493).

Este caso colonial deja claro que los desafíos morales implícitos en la sistematización contemporánea de la violencia tienen una historia más extensa[2]. Matar por una lógica de sistema reduce en gran medida, si no elimina, la posibilidad de conexión interpersonal, o incluso de reconocimiento. La persona cosificada que recibe la fuerza letal tiene poca o ninguna agencia en el proceso de selección de objetivos; no hay forma de saber cómo se desagrupan y reagrupan sus datos en la producción de la categoría "objetivo enemigo". En estos sistemas se hacen inferencia y suposiciones que encasillan categorías como la "enemistad" en términos discretos. Esta lógica se deriva de la ambición totalitaria de conocer al enemigo basado en la clasificación de datos y la tabulación cruzada. Las políticas nazis se caracterizaron por un proceso de cosificación y deshumanización. La clasificación sistemática de los seres humanos para su eliminación masiva rompió la premisa misma de las relaciones humanas: la de ser considerado como un individuo: un sujeto, no un objeto. Hanna Arendt contaba que, bajo el cuerpo de las SS, "la bestialidad dio paso a una destrucción absolutamente fría y

[2] Examinado en el punto anterior.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 237-257
https://doi.org/10.18543/djhr.3083 • http://djhr.revistas.deusto.es/

sistemática de los cuerpos humanos; calculada para destruir la dignidad humana y matar a cualquier individuo de los encarcelados en los campos" (Arendt 2018, 585). Los reclusos se convertían en objetos, clasificados en base a un sistema de identificación "según el cual cada prisionero tenía un identificador cosido a su uniforme" sobre el que se colocaba un "triángulo de clasificación" que indicaba por color si esa persona estaba categorizada como preso político, testigo de Jehová, prostituta u otro "asocial", homosexual, criminal, judío, etc. (Lifton 2016). Aprovechando las nuevas tecnologías para matar a distancia, tanto física como socialmente, y de este modo evitar los "horrores de la matanza cara a cara", la violencia de la Alemania nazi ofrece un claro ejemplo del desafío moral de la violencia sistemática desapasionada (Anders 1962).

Quizás estos ejemplos puedan no ser relevantes para los debates actuales sobre los peligros de las armas autónomas siendo demasiado exagerados. Pero es importante remarcar que no se pretende establecer una equivalencia moral entre las prácticas genocidas de la Segunda Guerra Mundial y el uso de la matanza autónoma a distancia. Lo que estos ejemplos históricos muestran es la matanza sistemática en su forma más patológica. El análisis de estos casos nos permite reconocer mejor las características problemáticas de la sistematización como proceso que opera en otros lugares, aunque en grados mucho menos severos. La sistematización de la violencia es un espectro que va desde lo rutinario y no problemático hasta lo asesino y genocida. En el medio no hay una serie de ejemplos de asesinatos sistemáticos que no son genocidas, pero siguen estando radicalmente en tensión con las normas morales prevalecientes.

Un ejemplo de ello es la estrategia de Estados Unidos en Vietnam durante la década de 1960, influenciada por la doctrina militar bajo Robert McNamara, centrada en un enfoque altamente cuantitativo y computacional conocido como "tecnoguerra". McNamara y su equipo creían que, con suficientes datos, la guerra podría ser completamente racional y controlable. Este enfoque llevó a un énfasis excesivo en evaluaciones de costo-beneficio y al "recuento de cadáveres" como métrica de éxito, resultando en la orden de matar a tantos enemigos sospechosos como fuera posible. La amplia y a menudo imprecisa clasificación del enemigo facilitó la comisión de numerosas atrocidades. Se estima que, entre 1,1 y 3,8 millones de vietnamitas, tanto civiles como combatientes, murieron violentamente (Anders 1962). Este sistema incentivó la muerte y corrompió los valores morales. Estos riesgos persisten en la guerra algorítmica actual, amenazando tanto la moralidad de los objetivos como de quienes ejecutan la violencia.

De todo esto surge la siguiente pregunta: ¿cómo pudieron hacerlo? ¿Cómo pudieron todos ser partícipes de un sistema de asesinatos producidos en masa? La respuesta a esta pregunta nos ayudará a comprender mejor los peligros presentes y futuros de la matanza autónoma. Frente a ciertas atrocidades, se reconoce que la hostilidad "ha estado arraigada históricamente e inducida situacionalmente por las condiciones bajo las cuales se debilitan las inhibiciones morales habituales contra la violencia" (Kelman 1973, 25-61). Se identifica entonces que la autorización, la rutinización y la deshumanización son importantes contribuyentes a este debilitamiento de la restricción moral.

La autorización proporciona el sustrato necesario para las transgresiones sancionadas a gran escala. Cuando un agente legítimo y autoritario ordena explícitamente o aprueba actos de violencia, la disposición del agente a tolerarlo aumenta considerablemente. Mediante la autorización, el control se entrega a agentes autorizados vinculados a objetivos más amplios, a menudo abstractos, que van más allá de las reglas de la moralidad común (Anders 1962). Para aquellos, a los que se les permite hacer uso de la violencia de manera efectiva, la responsabilidad se ve dispersa en manos de las autoridades centrales, que a su vez ceden su autoridad a cargos aún más altos. Lo que causa que se separe la parte cognitiva de los afectos y la amoralidad personal de una apelación racionalizada a la violencia dominante.

Por otro lado, la rutinización erosiona las restricciones morales, mientras que la autorización anula las preocupaciones morales que de otro modo existirían; los procesos de rutinización limitan los puntos en los que las preocupaciones morales pueden surgir. La rutinización cumple dos funciones: la primera, reducir la necesidad de tomar decisiones, minimizando las ocasiones en las que pueden surgir cuestiones morales; y, la segunda, hace que sea más fácil evitar las implicaciones que conlleva la acción, pues el actor se centra en los detalles más que en el significado de la tarea en cuestión. A su vez, los procesos de deshumanización privan a las víctimas de su condición humana. "En la medida en que las víctimas son deshumanizadas, los principios de moralidad ya no se aplican a ellas y las restricciones morales contra el asesinato se superan más fácilmente" (Anders 1962, 1). A su vez, estos mismos procesos de degradación del estatus moral de la víctima también pueden deshumanizar a los perpetradores.

> A través de su obediencia incondicional a la autoridad y a través de la rutinización de su trabajo, se le priva del albedrío personal. No es un actor independiente que emite juicios y decisiones sobre la

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 237-257
https://doi.org/10.18543/djhr.3083 • http://djhr.revistas.deusto.es/

> base de sus propios valores y de la evaluación de las consecuencias. Más bien, se deja zarandear por fuerzas externas. Se enajena dentro de su tarea (Anders 1962, 1).

De lo dicho anteriormente podemos enfatizar que la sistematización como problema no es específica de ninguna tecnología o modo de guerra. Podemos observar el programa de aviones no tripulados armados de EE.UU. para una ilustración más reciente de los efectos problemáticos de la autorización, la rutinización y la deshumanización. Dentro de este programa, los aviones no tripulados armados eran parte de una "red flexible" de capacidades que abarcaban una distancia global y se entrelazaban mediante matrices de transmisión de datos (Schultz 2021). A partir de esta tecnología de integración de sistemas surgieron varios desafíos morales, particularmente en el contexto de los asesinatos selectivos. Inicialmente este programa se justificó como una respuesta necesaria al terrorismo, derivando en un deterioro de los estándares de selección de objetivos a medida que los asesinatos con drones se volvieron más rutinarios (BBC News 2013). Durante la candidatura de Obama se hizo referencia a la naturaleza sistemática de los asesinatos con aviones no tripulados de EE.UU. y a la perdida moral que incentivó:

> El problema con el programa de drones... es que empieza a darte la ilusión de que eso no era la guerra... la maquinaria comenzó a volverse demasiado fácil de utilizar, y tuve que imponer internamente un conjunto sustancial de reformas en el proceso para dar un paso atrás y recordar a todos los involucrados que esto no es una práctica de tiro (CBS News 2020).

La deshumanización también afectó al programa de aviones no tripulados de EE.UU., agravada por la naturaleza basada en datos de los asesinatos: los objetivos a veces se comparan como si fueran malezas o plagas (Pilkington 2017). Las opiniones internas de la comunidad hacia los perseguidos por drones armados eran: "no tienen derechos. No tienen dignidad. No tienen humanidad para sí mismos. Solo son un 'selector de objetivos' para un analista. Al final llegas a un punto en el ciclo de vida del objetivo en el que simplemente lo sigues como si fuera un objeto, ni siquiera te refieres a él por su nombre real. Esta práctica, contribuye a deshumanizar a la gente incluso antes de que te hayas encontrado con la pregunta moral de ¿es este un asesinato legítimo o no?" (Scahill 2015, 93-110).

Tras una extensa revisión de los diferentes tipos de asesinato sistemático a lo largo de varios periodos históricos, que varían

considerablemente tanto en los medios como en los fines de la violencia, se pueden identificar puntos en común. Aunque la violencia sistemática no es intrínsecamente problemática, plantea desafíos morales ineludibles, especialmente en casos de sistematización intensificada. Esto incluye la erosión del estatus moral tanto de quienes ejercen como de quienes reciben la violencia, una pérdida que puede impactar negativamente en la moderación durante la guerra; un riesgo que persiste en el contexto actual de la violencia autónoma sistemática.

4. **¿Hay posibilidad de retroceder?**

La carrera armamentística en torno a las armas autónomas se ha intensificado en los últimos años, impulsada por avances tecnológicos y el deseo de mantener una ventaja militar. A pesar de los profundos dilemas éticos y las amenazas a la estabilidad global, el retroceso en esta carrera parece prácticamente imposible. Esto se debe en gran medida a la disposición de ciertos países a priorizar la superioridad militar sobre la preservación de valores morales y la estabilidad internacional. A continuación, se exploran las razones detrás de esta dinámica y sus implicaciones.

Las armas autónomas ofrecen una promesa tentadora de superioridad militar. La capacidad de desplegar sistemas de combate que operan con velocidad, precisión y sin la necesidad de intervención humana directa proporciona una ventaja estratégica significativa. Países como Estados Unidos, China y Rusia están invirtiendo enormes recursos en el desarrollo de estas tecnologías, reconociendo que quien domine en este campo podría establecer una hegemonía militar (Scharre 2018).

El dilema de seguridad juega un papel crucial en la carrera armamentística. Si un país reduce su inversión en armas autónomas, corre el riesgo de quedar vulnerable frente a aquellos que continúan avanzando. Este temor de quedarse atrás crea una situación en la que ninguna nación quiere ser la primera en detener el desarrollo de estas tecnologías (Altmann y Sauer 2017). La percepción de que los adversarios potenciales podrían obtener una ventaja decisiva empuja a los países a seguir adelante, perpetuando la carrera.

La falta de confianza entre las naciones complica aún más la posibilidad de un acuerdo para detener o revertir el desarrollo de armas autónomas. Sin un marco internacional sólido y verificable, los países no pueden estar seguros de que sus rivales cumplirán con cualquier acuerdo de limitación. Esta desconfianza mutua fomenta un ciclo continuo de desarrollo y despliegue de armas autónomas (Payne 2021).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 237-257
https://doi.org/10.18543/djhr.3083 • http://djhr.revistas.deusto.es/

Establecer un régimen de control efectivo para las armas autónomas es extremadamente complicado. Las definiciones y límites sobre qué constituye un arma autónoma pueden ser ambiguas, y las tecnologías subyacentes, como la inteligencia artificial, son duales, es decir, pueden tener usos tanto civiles como militares. La regulación internacional enfrenta obstáculos significativos en la verificación del cumplimiento y la imposición de sanciones (Marchant et al. 2011).

Los complejos industriales-militares tienen un interés económico en el desarrollo continuo de nuevas tecnologías de armas, incluidas las autónomas. Las inversiones en investigación y desarrollo, así como las expectativas de lucrativos contratos de defensa, crean una fuerte inercia que impulsa la carrera armamentística. Estos actores suelen ejercer presión sobre los gobiernos para mantener y aumentar las inversiones en capacidades militares avanzadas (Singer 2009).

La posibilidad de retroceder en la carrera armamentística de las armas autónomas es extremadamente limitada debido a la combinación de la búsqueda de ventaja militar, el dilema de seguridad, la desconfianza internacional, los desafíos regulatorios, la presión de los complejos industriales-militares y la erosión de valores morales. Para mitigar los riesgos asociados es esencial un enfoque multilateral que fomente la cooperación internacional, la transparencia y la construcción de confianza; aunque lograrlo sea una tarea formidable en el actual clima geopolítico.

Conclusión

Los procesos de sistematización subsumen el juicio humano en el campo de batalla hasta un grado moralmente problemático. El derecho internacional humanitario valora el juicio humano precisamente porque, aunque históricamente ha demostrado ser un control insuficiente de la conducta individual, representa una salvaguarda ética esencial. El principio de discriminación, la prohibición de la perfidia y la protección de los prisioneros de guerra: estas reglas resisten a cualquier presión externa situacional que exista, es independiente de cualquier capricho del juicio de los combatientes. Sin embargo, estas reglas no deben cegarnos ante los peligros de lo contrario: formas frías y desapasionadas de violencia sistemática que erosionan el estatus moral de los objetivos humanos, así como el estatus de aquellos que participan dentro del propio sistema.

El argumento de que los LAWS pueden ser agentes más éticos en la guerra solo puede sostenerse si pensamos en la guerra como una

actividad en gran medida procesal y centrada en el proceso en la que las líneas morales son relativamente fáciles de identificar y lo suficientemente firmes como para soportar la incertidumbre y la ambigüedad. Esto puede ser un ideal, pero no lo es, y probablemente nunca será una realidad. La guerra está dividida por una complejidad que impide la certeza; y, por extensión, la aplicación fluida y fiable de la violencia sistemática a los sujetos que son un objetivo. Actuar como si esta no fuera la realidad, imponer la violencia sistemática en entornos estructuralmente inadecuados para tal enfoque, es cortejar un daño moral previsible y ruinoso.

Las LAWS y los asesinatos infundidos por la IA, sistematizan la violencia en el sentido más literal. El sistema proporciona la organización, la función optimizada, el distanciamiento y el vacío moral necesarios para expandir los modos de matar en lugar de fomentar la restricción. Esto no es un genocidio ni una limpieza étnica ni ninguna de las otras formas de asesinato sistemático histórico examinadas en este artículo. La violencia de las LAWS no se acerca moralmente a la matanza masiva que marcó gran parte del siglo XX. Pero sí se observa, sin embargo, un eco del pasado problemático en los procesos autónomos de hoy: un conjunto implícito de condiciones que podrían facilitar la infracción moral en el uso de la violencia letal en la guerra.

Retroceder en la carrera armamentística de las armas autónomas letales es poco realista en la actualidad, ya que requeriría un esfuerzo coordinado y multifacético que involucre voluntad política, cooperación internacional, desarrollo de regulaciones y tratados, y la consideración de las implicaciones éticas y humanitarias. La historia de los tratados de control de armas ofrece precedentes de éxito, pero también destaca los desafíos que se deben superar para lograr un control efectivo de las LAWS.

> Nada es tan peligroso en el estudio de la guerra como permitir que las máximas se conviertan en un sustituto del juicio (Corbett 2008, p.167).

Referencias

Altmann, Jürgen y Frank Sauer. 2017. «Autonomous weapon systems and strategic stability». *Survival* 59 (5): 117-142.

Anders, Günther. 1962. *Burning conscience.* Nueva York: Monthly Review Press

Ansorge, Josef T. 2016. *Identify and sort: How digital power changed world politics*. Oxford: Oxford University Press.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 237-257
https://doi.org/10.18543/djhr.3083 • http://djhr.revistas.deusto.es/

Arendt, Hannah. 2018. *The origins of totalitarianism.* Galway: Penguin.
Asaro, Peter. 2012. «On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making». *International Review of the Red Cross* 94 (886): 687–709. doi:10.1017/S1816383112000768
BBC News. 2013. *US Drone Programme: 'Strict, fair and accountable' - Kerry.* Acceso el 28 de mayo de 2024. https://www.bbc.com/news/av/world-radio-and-tv-22690918/us-droneprogramme-strict-fair-and-accountable-kerry.
Best, Geoffrey Francis Andrew. 1980. *Humanity in warfare. The modern history of the international law of armed conflicts.* Oxford: Routledge.
Bode, Ingvild y Thomas Watts. 2021. *Meaning-less human control: Lessons from air defence systems on meaningful human control for the debate on AWS*. Odense: University of Hertfordshire.
CBS News. 2020. *Barack Obama: The 2020 60 minutes interview*. Acceso el 16 de noviembre de 2023. https://www.cbsnews.com/video/barack-obama-60-minutes-2020-11-15.
Cheney-Lippold, John. 2019. «Accidents happen». *Deleted Journal* 86 (2): 513-35.
Congressional Research Service. 2022. «Defense Primer: command and control», *In Focus*, actualizado el 14 de Noviembre.
Corbett, Julian. 2008. *Some principles of maritime strategy.* Milton Park: Routledge.
Crootof, Rebecca. 2019. «Regulating new weapons technology». En The impact of emerging technologies in the Law of armed conflict, editado por Ronald T.P. Alcala y Eric Talbot, 3-36. Oxford: Oxford University Press eBooks. doi.org/10.1093/oso/9780190915322.003.0001.
Cummings, Mary. 2004. «Automation bias in intelligent time critical decision support systems». *American Institute of Aeronautics and Astronautics*, junio. doi.org/10.2514/6.2004-6313.
Ekelhof, Merel. 2019. «Moving beyond semantics on autonomous weapons: Meaningful human control in operation». *Global Policy* 10 (3): 343-48. doi.org/10.1111/1758-5899.12665.
Emery, John R. 2016. «Review: Anthony F. Lang Jr., Cian o'Driscoll y John Williams, eds. Just War: Authority, Tradition, and Practice». *European Review of International Studies* 3 (1): 143-47. doi.org/10.3224/eris.v3i1.26023.
Freedman, Lawrence. 2017. *The future of war: A history*. Londres: Allen Lane.
Glover, Jonathan. 2000. *Humanity: A moral history of the Twentieth Century*. Londres: Yale University Press. http://ci.nii.ac.jp/ncid/BA44910039.
Heyns, Christof. 2017. «Autonomous weapons in armed conflict and the right to a dignified life: an African perspective». *South African Journal of Human Rights* 33 (1): 46–71. doi.org/10.1080/02587203.2017.1303903
Keegan, John. 1994. *A history of warfare.* Londres: Pimlico.
Kelman, Herbert G. 1973. «Violence without moral restraint: Reflections on the dehumanization of victims and victimizers». *Journal of Social Issues* 29 (4): 25-61. doi.org/10.1111/j.1540-4560.1973.tb00102.x.

Knight, Will. 2021. «The Pentagon inches toward letting AI control weapons». *WIRED*. Acceso el 10 de mayo de 2023. https://www.wired.com/story/pentagon-inches-toward-letting-ai-control-weapons/.

LeMoncheck, Linda. 1985. *Dehumanizing women: Treating persons as sex objects*. Londres; Rowman & Allanheld.

Lifton, Robert J. 2016. «The genocidal mentality». *Tikkun* 31 (3): 32-33. doi.org/10.1215/08879982-3628248.

Marchant, Gary E., Braden R. Allenby, y Joseph R. Herkert. 2011. *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*. Dordrecht: Springer.

Mitchell, Melanie. 2019. *Artificial Intelligence: A guide for thinking humans*. Nueva York: Farrar, Straus y Giroux.

Münkler, Herfried. 2004. *Die neuen kriege*. Hamburgo: Rowohlt. http://ci.nii.ac.jp/ncid/BB06266147.

Nussbaum, Martha C. 1995. «Objectification». *Philosophy and Public Affairs* 24 (4): 249-91. doi.org/10.1111/j.1088-4963.1995.tb00032.x.

Pakenham, Thomas. 1979. *The Boer war*. Barcelona: Futura Book. http://ci.nii.ac.jp/ncid/BA86489462.

Payne, Kenneth. 2021. *The dawn of artificially intelligent conflict*. Londres: Hurst Publishers.

Phillips-Levine, Trevor, Michael Kanaan, Dylan Phillips-Levine, Walter D. Mills y Noah Spataro. 2022. «Weak human, strong force: applying advanced chess to military AI». *War on the Rocks*. Acceso el 7 de julio de 2023. https://warontherocks.com/2022/07/weak-human-strong-force-applying-advanced-chess-to-military-ai.

Pilkington, Ed. 2017. «Life as a drone operator: 'Ever step on ants and never give it another thought?'» *The Guardian*, 14 de julio. Acceso el 4 de julio de 2024. https://www.theguardian.com/world/2015/nov/18/life-as-a-drone-pilot-creech-air-force-base-nevada.

Riesen, Erich. 2022. «The moral case for the development and use of autonomous weapon systems». *Journal of Military Ethics* 21 (2): 132-50. doi.org/10.1080/15027570.2022.2124022.

Scahill, Jeremy. 2015. «Leaked military documents expose the inner workings of Obama's drone wars». *The Intercept.* 21 de octubre. Acceso el 21 de octubre de 2024. https://www.theintercept.com/drone-papers/the-assassination-complex.

Scharre, Paul. 2018. *Army of none: Autonomous weapons and the future of war*. Nueva York: W. W. Norton & Company.

Schultz, Timothy P. 2021. *Remote warfare: A new architecture of air power*. Cambridge: Cambridge University Press. doi.org/10.1017/9781108985024.003.

Schwarz, Elke. 2021. «Autonomous weapons systems, artificial intelligence, and the problem of meaningful human control». *Philosophical Journal of Conflict and Violence* 5 (1): 53-72. doi.org/10.22618/tp.pjcv.20215.1.139004.

Singer, P. W. 2009. *Wired for war: The robotics revolution and conflict in the 21st century*. Nueva York: Penguin Books.

Umbrello, Steven, Phil Torres, y Angelo F. de Bellis. 2019. «The future of war: could lethal autonomous weapons make conflict more ethical?». *AI & Society* 35 (1): 273-82. doi.org/10.1007/s00146-019-00879-x.

Walker, Paddy. 2021. «Leadership challenges from the deployment of lethal autonomous weapon systems: How erosion of human supervision over lethal engagement will impact how commanders exercise leadership». *The RUSI Journal* 166 (1): 10–21. doi:10.1080/03071847.2021.1915702.

Williams, John. 2021. «Locating LAWS: Lethal autonomous weapons, epistemic space, and 'meaningful human' control». *Journal of Global Security Studies* 6 (4): 1-18. doi.org/10.1093/jogss/ogab015.

Los MASC como derecho humano para optar por otra forma de justicia y la IA como vía para facilitar su efectividad

The ADR as a human right to choose another form of justice and AI as a way to facilitate its effectiveness

Ana María Vall Rius
CUNEF Universidad. España
ana.vall@cunef.edu
ORCiD: https://orcid.org/0000-0002-5649-8173

https://doi.org/10.18543/djhr.3196

Fecha de recepción: 29.05.2024
Fecha de aceptación: 22.11.2024
Fecha de publicación en línea: diciembre de 2024

Cómo citar / Citation: Vall, Ana Mª. 2024. «Los MASC como derecho humano para optar por otra forma de justicia y la IA como vía para garantizar su efectividad». *Deusto Journal of Human Rights*, n. 14: 259-285. https://doi.org/10.18543/djhr.3196

Sumario: Introducción. 1. Concepto y tipología de los MASC. 2. Los MASC y su papel en los diversos ámbitos de las relaciones interpersonales. 3. Los recursos de la inteligencia artificial como facilitadores de los MASC. Conclusiones y recomendaciones. Referencias.

Resumen: El acrónimo MASC (Medios Adecuados de Solución de Controversias) engloba un conjunto de metodologías de gestión pacífica de conflictos que suponen una nueva forma de entender la justicia, en la cual los propios ciudadanos se implican directamente en la búsqueda de soluciones constructivas. Por ello los compromisos que se alcanzan, además de ser consensuados, responden mejor a las necesidades y circunstancias de cada persona, de cada caso y situación. Numerosos expertos en la materia entienden que, poder optar por la aplicación efectiva de alguno de estos métodos para gestionar las discrepancias y conflictos interpersonales, es un derecho que ha de estar al alcance y a la disposición de todas las personas sin ningún tipo de discriminación u obstáculo insalvable, al igual que el derecho de acceso a la Justicia, ya que estos métodos ofrecen un nuevo sentido a la vieja expresión de "hacer justicia". Por ello es necesario contar con sistemas de la llamada Inteligencia Artificial (IA) que faciliten que este nuevo paradigma pueda hacerse realidad también en situaciones en las cuales, por diversas circunstancias, el encuentro personal presencial entre las partes en conflicto no es factible o no es deseable.

Palabras clave: MASC, justicia, derechos humanos, conflictos, justicia restaurativa, soluciones, consenso, acuerdos.

Abstract: The acronym ADR (Appropriate Dispute Resolution) or in Spanish MASC (Appropriate Methods of Dispute Resolution) encompasses a set of peaceful conflict management methodologies that represent a new way of understanding justice, in which citizens themselves are directly involved in the search for constructive solutions. For this reason, the commitments that are reached, in addition to being consensual, respond better to the needs and circumstances of each person, each case and situation. Many experts in the field understand that being able to choose the effective application of any of these methods to manage discrepancies and interpersonal conflicts is a right that must be within the reach and disposal of all people without any type of discrimination or insurmountable obstacle, at the same time, as well as the right of access to Justice, since these methods offer a new meaning to the old expression of "doing justice." For this reason, it is necessary to have so-called Artificial Intelligence (AI) systems that facilitate this new paradigm also becoming a reality in situations in which, due to various circumstances, a personal meeting between the parties in conflict is not feasible or is not desirable.

Keywords: ADR, justice, human rights, conflicts, restorative justice, solutions, consensus, agreements.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

Introducción

Tanto el derecho internacional[1], como la normativa interna de los distintos países de cultura jurídica occidental, entre ellos el nuestro, establecen, fomentan y regulan la posibilidad de optar por los MASC (Medios Adecuados de Solución de Controversias) como un recurso voluntario y a la vez un derecho de los ciudadanos en general, aplicable en diversos ámbitos[2]. Se trata de un derecho humano, asimilable al

[1] *Declaration of the Ministers of Justice of the Council of Europe Member Estates on the role of restorative justice in Criminal Matters* (con occasion de la Conferencia de Ministros de Justicia del Consejo de Europa *Crime and Criminal Justice. The role of restorative justice in Europe*". Venecia 13 y 14 de diciembre de 2021, conocida también como "Declaración de Venecia".

- Recomendación CM/Rec. (2018) 8 del Comité de Ministros de los Estados miembros en materia de justicia restaurativa penal.
- *"Rebooting" the Mediation Directive: Assessing the limited impact of its implementation and proposing measures to increase the number of mediations in the EU* (enero de 2014).
- Directiva 2012/29/UE del Parlamento Europeo y del Consejo de 25 de octubre de 2012, por la que se establecen normas mínimas sobre los derechos, el apoyo y la protección de las víctimas de delitos y por la que se sustituye la Decisión marco 2001/220/JAI del Consejo.
- Directiva 2008/52/CE del Parlamento Europeo y del Consejo, de 21 de mayo de 2008, sobre ciertos aspectos de la mediación en asuntos civiles y mercantiles. DOUEL n. 136, de 24 de mayo de 2008. Véase: https://www.boe.es/doue/2008/136/L00003-00008.pdf
- Libro Verde sobre las modalidades y alternativas de solución de conflictos en el ámbito del derecho civil y mercantil, COM 2002/0196. Presentado por la Comisión.
- Manual sobre Programas de Justicia Restaurativa de 2006 de las Naciones Unidas (UNODC 2006).
- Principios básicos sobre la utilización de programas de justicia restaurativa en materia penal de las Naciones Unidas (ECOSOC Res. 2002/12)
- Recomendación n. R (98) 1, del Comité de Ministros a los Estados Miembros sobre la Mediación Familiar. (Aprobada por el Consejo de Ministros del 21 de enero de 1998, a partir de la 616 reunión de los Delegados de los Ministros).

[2] Por ejemplo, Italia, donde el vigente Decreto-Ley n. 69/2013 (convertido en ley mediante *la Legge di Conversione 9 agosto 2013 n. 98*) declara que las partes deben someterse a mediación con carácter obligatorio, asistidas de sus respectivos letrados cuando el conflicto tenga por objeto alguna de las materia contempladas en el art. 5 (derechos reales, división o partición de herencia, arrendamiento de vivienda o negocio, indemnización de daños derivados de responsabilidad médico-sanitaria, seguros, banca y contratos financieros). Esta obligación de someterse a mediación se limita al deber de asistir y celebrar una primera sesión de mediación y a no seguir el proceso de mediación si una o todas las partes no lo desean. Dicha obligatoriedad ya se había establecido en Italia previamente mediante la promulgación del Decreto-Legislativo n. 28/2010 convertido en la *Legge 18 giugno de 2009 n. 69 in materia di mediazione finalizzata allá conciliazione*

reconocido derecho humano de acceso a la justicia recogido en destacados textos jurídicos[3]. Consiste en facilitar que sea posible escoger la opción de un nuevo modelo de justicia, que debe estar a la disposición de todas las personas para facilitar la gestión pacífica y eficiente de sus controversias sin tener que recurrir necesariamente a la vía judicial contenciosa. Acudir al sistema tradicional de Justicia supone ceder a terceros el poder de decidir sobre cuestiones que son importantes en la vida de las personas y que, por el principio dispositivo, predominante en el derecho privado, pueden gestionar y resolver los mismos protagonistas, partiendo de la evidencia de que las propias personas afectadas conocen mejor que nadie su realidad y los múltiples factores que condicionan sus circunstancias. Además, muchas de las cuestiones que se dirimen en los juzgados y tribunales no tienen un cariz ni total ni parcialmente jurídico, sino que, en un elevado porcentaje, se trata de discrepancias básicamente personales y relacionales, nacidas de necesidades insatisfechas, que generan consecuencias legales a las que convendría dar una respuesta integral y no meramente jurídica.

Si el objetivo es resolver realmente estos conflictos y evitar que acaben convirtiéndose en endémicos, habrá que buscar soluciones holísticas que permitan responder adecuadamente al origen personal del conflicto en sus múltiples facetas, sin olvidar sus consecuencias jurídicas.

La metodología utilizada en este trabajo se basa en el análisis diferenciado de los distintos elementos, características y valores fundamentales de los que se parte (los MASC, las diferencias entre los distintos ámbitos de aplicación de estos métodos y la IA) para ponerlos finalmente en relación destacando las oportunidades y ventajas de esta conexión desde una visión, no meramente teórica, sino desde la práctica y la experiencia profesional de más de 20 años.

1. **Concepto y tipología de los MASC**

El acrónimo MASC engloba los distintos medios que facilitan esta implicación directa de las personas protagonistas en la gestión y

delle controversia civili e commezialle. Pero dicho Decreto fue declarado inconstitucional por la sentencia 272/2012 del Tribunal Constitucional italiano, que declaró inconstitucional el Decreto-Legislativo n. 28/2010, por razones de procedimiento.

[3] Arts. 8 y 10 de la Declaración Universal de los Derechos Humanos, proclamada por la Asamblea General de las Naciones Unidas, en París el 10 de diciembre de 1948 (Resolución 217 A III).

Art. 24 de la Constitución Española de 1978, según el cual todas las personas tienen derecho a la tutela judicial efectiva.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

resolución colaborativa de los conflictos y discrepancias interpersonales. Estos métodos permiten que las personas que protagonizan una controversia puedan ser también actores principales en el proceso de toma de decisiones, que está encaminado a encontrar soluciones consensuadas más personalizadas y que den una mayor satisfacción a las necesidades concretas del caso. Un tercero profesional les ayudará en este proceso de búsqueda de soluciones satisfactorias para todos. Este tercero no les impone nada, sino que contribuye a través de distintas técnicas y herramientas a que alcancen soluciones óptimas a través del consenso y de la negociación asistida. Se parte de la premisa de que son las personas que viven un conflicto las que conocen mejor que nadie su propia realidad, sus intereses, sus necesidades y circunstancias. Por ello la mejor solución es la que puede emanar de ellos mismos. Además, al hacerse conscientes de que son ellos quienes han construido esta respuesta común, es mucho más fácil que después la pongan en práctica y que la cumplan según lo acordado. Desde otra perspectiva, estos métodos no parten tanto de la idea de confrontar o de demostrar quién tiene razón o quien pueda tener algún tipo de culpa, sino que el objetivo principal es el de buscar soluciones que puedan ser útiles para todos y que estén basadas en la colaboración y el consenso. Estamos, por tanto, ante una solución mucho más sólida, realista, eficaz y menos controvertida que la imposición de una sentencia judicial.

Los diversos MASC tienen elementos comunes como es la implicación directa de los protagonistas en la búsqueda de resultados consensuados y que parten de la plena voluntariedad de iniciar o de rechazar uno de estos métodos que, además, pueden abandonar en todo momento. A la vez cada uno de los MASC utiliza estrategias propias y tiene diferentes *modus operandi.*

El método más utilizado, geográficamente más expandido y consolidado actualmente es la mediación, a pesar de que como señala Martín Diz (2020, 1), todavía es un concepto que puede sonar extraño. En la mediación el tercero mediador crea un espacio en el que las partes pueden dialogar, tratar todo aquello que les preocupa y exponer cuáles son sus necesidades e intereses y aquellos objetivos que pretenden alcanzar a través de la mediación. La persona mediadora les facilita esa comunicación bidireccional, les ayuda a identificar sus necesidades e intereses, tanto individuales como comunes, y motiva la elaboración y el intercambio de propuestas, a través de las cuales pueden alcanzar el consenso y confeccionar una propuesta común. En la mediación facilitativa, que es la más común

en nuestro país, la persona mediadora potencia el papel protagonista de los participantes en la mediación y propicia la búsqueda de soluciones consensuadas, sin realizar ningún tipo de proposición o sugerencia, ya que está convencido de que las mejores propuestas son las que emanan de ellos mismos.

La negociación también se contempla como uno de los métodos MASC y posiblemente sea el más utilizado por parte de los letrados (negociación entre abogados) o incluso entre las personas protagonistas de una controversia que directamente, sin necesidad de acudir a un tercero, tratan de encontrar una solución por ellos mismos, a través de una negociación directa para alcanzar un pacto consensuado que les permita superar sus discrepancias.

Otro método bastante usual es la conciliación privada, en la cual el tercero igualmente facilita la comunicación, pero su intervención en el desarrollo del proceso es más intensa al tener la posibilidad de presentar a las partes propuestas de solución no vinculantes que las partes pueden aceptar o rechazar, o bien modificar total o parcialmente, de forma libre y voluntaria.

El Proyecto de Ley Orgánica de medidas en materia de eficiencia del Servicio Público de Justicia, fue publicado en el Boletín Oficial de las Cortes del 22 de marzo de 2024 y aprobado por el Congreso de los Diputados el 14 de noviembre de 2024 (algunas informaciones apuntan a que podría producirse su aprobación antes de finalizar el año). Esta futura Ley Orgánica incluye, además de la mediación y la negociación, otros métodos como la conciliación privada, ya mencionada, la oferta vinculante, el derecho colaborativo y la opinión de un experto independiente. Además, su art. 2 entiende como medio adecuado de solución de controversias "cualquier tipo de actividad negociadora, reconocida en esta u otras leyes estatales o autonómicas, a la que las partes de un conflicto acuden de buena fe con el objeto de encontrar una solución extrajudicial al mismo, ya sea por sí mismas o con la intervención de una tercera persona neutral". Como señaló el presidente del Consejo de la Abogacía Española, Salvador González, refiriéndose a esta futura ley: "Era necesaria, es mejorable, pero nos preocupa tanto la ley como su puesta en funcionamiento, que se destinen los recursos necesarios…"[4].

El Proyecto contempla también un método menos conocido y poco utilizado todavía en nuestro país: la Oferta vinculante confidencial,

[4] Véase: https://www.abogacia.es/actualidad/noticias/el-congreso-aprueba-la-ley-organica-de-eficiencia-del-servicio-publico-de-justicia/

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

regulada en su artículo 17 de la forma siguiente: "Cualquier persona que, con ánimo de dar solución a una controversia, formule una oferta vinculante a la otra parte, queda obligada a cumplir la obligación que asume, una vez que la parte a la que va dirigida la acepta expresamente. Dicha aceptación tendrá carácter irrevocable". Se trata de un método que puede ser muy útil y agilizar soluciones pactadas, especialmente en el campo de los negocios y de las transacciones comerciales. En la práctica, en la mayor parte de las ocasiones esta oferta se presentará de forma electrónica a través de un sistema que permita certificar su existencia y autenticidad.

En el artículo 18 se contempla la opinión de un experto independiente: "Las partes, con objeto de resolver una controversia, podrán designar de mutuo acuerdo a una persona experta independiente para que emita una opinión no vinculante respecto a la materia objeto de conflicto. Las partes estarán obligadas a entregar a la persona experta toda la información y pruebas de que dispongan sobre el objeto controvertido". Según el mismo artículo, dicho dictamen podrá versar sobre cuestiones jurídicas o sobre cualquier otro aspecto técnico y tendrá carácter confidencial. También esta fórmula puede ser muy eficiente para superar las discrepancias de tipo técnico surgidas en diversos ámbitos de las relaciones interpersonales y especialmente en aquellas obligaciones derivadas de los contratos de obra y de servicios.

Además, es muy relevante que dicho Proyecto de Ley establezca en su artículo 5 el requisito de procedibilidad con carácter general, en materias civiles que va a suponer una importante novedad y un impulso considerable para la utilización de los MASC en nuestro país. El requisito de procedibilidad implica que, para que el órgano judicial correspondiente admita la demanda se considerará requisito imprescindible para su admisibilidad el haber acudido previamente a algún medio adecuado de solución de controversias como los que hemos visto y que están previstos en su artículo 2: mediación, negociación, conciliación, oferta vinculante confidencial, la opinión de un experto independiente "o cualquier otro tipo de actividad negociadora tipificada en esta u otras leyes estatales o autonómicas…" (art. 5).

Por tanto, deberá acreditarse, tal como establece el Proyecto, que, de forma previa a la interposición de la demanda contenciosa, se ha intentado solucionar de manera consensuada la controversia mediante alguno de estos medios autocompositivos.

2. Los MASC y su papel en los diversos ámbitos de las relaciones interpersonales

En el ámbito del derecho privado civil, mercantil, societario y empresarial, se contempla la disponibilidad de muchas cuestiones, como las referidas a la negociación, interpretación y ejecución de las cláusulas y compromisos contractuales, división de cosa común, discrepancias en el seno de las sociedades, desacuerdos entre los integrantes de los órganos directivos y de gestión, etc. Esta prioridad que nuestro derecho otorga, en este ámbito, al libre albedrio consensuado, refleja el amplio reconocimiento del poder de decidir de las partes, cuya preeminencia viene establecida por el propio derecho frente a la regulación que establece la ley. La norma queda como un recurso subsidiario al que recurrir cuando las partes no alcanzan un acuerdo sobre algún punto en discrepancia.

Incluso, según establece el art. 19-1 de la Ley 1/2000 de Enjuiciamiento Civil, una vez ya iniciado el proceso judicial civil, siempre que la ley no lo prohíba expresamente o introduzca límites por razón del interés general o en beneficio de tercero, los litigantes podrán disponer del objeto del juicio "y podrán renunciar, desistir del juicio, allanarse, someterse a mediación o arbitraje y transigir sobre lo que sea objeto del mismo".

También en otros ámbitos diferentes al derecho privado, como en el derecho de familia con afectación de menores o de otros miembros especialmente vulnerables, las personas responsables pueden tomar decisiones y presentar propuestas de solución que repercuten en las propias relaciones familiares. Así lo establece el art. 90-2 del Código Civil[5] y el 91 del mismo Código[6], entre otras disposiciones, que dan preferencia a las decisiones que han sido acordadas por las partes.

[5] "2. Los acuerdos de los cónyuges adoptados para regular las consecuencias de la nulidad, separación y divorcio presentados ante el órgano judicial serán aprobados por el juez salvo si son dañosos para los hijos o gravemente perjudiciales para uno de los cónyuges".

[6] "En las sentencias de nulidad, separación o divorcio, o en ejecución de las mismas, la autoridad judicial, en defecto de acuerdo de los cónyuges o en caso de no aprobación del mismo, determinará conforme a lo establecido en los artículos siguientes las medidas que hayan de sustituir a las ya adoptadas con anterioridad en relación con los hijos, la vivienda familiar, el destino de los animales de compañía, las cargas del matrimonio, liquidación del régimen económico y las cautelas o garantías respectivas, estableciendo las que procedan si para alguno de estos conceptos no se hubiera adoptado ninguna. Estas medidas podrán ser modificadas cuando se alteren sustancialmente las circunstancias".

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

Estas propuestas responderán mucho mejor a las necesidades de las personas afectadas, ya que emanan de quienes viven en directo esa realidad y son realmente conocedores del caso concreto, de su complejidad y de sus circunstancias. Los órganos judiciales validarán la propuesta presentada, si los bienes jurídicos afectados y en especial el interés superior del menor o de las personas más vulnerables, quedan adecuadamente preservados. En la práctica aquellos casos de discrepancias familiares en los que se alcanza un consenso y las mismas partes elaboran su propia propuesta, ésta se presenta en forma de convenio regulador o en el documento formal que corresponda y el Juez suele aprobarlo en la gran mayoría de los supuestos valorando la implicación directa de las partes y la mayor adecuación de la respuesta a las concretas necesidades de cada caso, ya que dicha propuesta emana de los integrantes del propio núcleo familiar[7].

Otro campo diferente al derecho privado o al derecho de familia es el correspondiente al derecho penal en el cual el ámbito de decisión de las partes es más reducido, al tratarse de normas mayoritariamente imperativas. Aunque esto no ha de significar necesariamente que la voluntad de las personas involucradas, voluntaria o involuntariamente, en un hecho delictivo quede absolutamente anulada y que esa capacidad de tomar decisiones como derecho humano deba ser

[7] Numerosas sentencias avalan esta confirmación por parte del Juez de los compromisos y acuerdos que las partes han consensuado, reforzando el valor de alcanzar acuerdos y consensos, a través de la mediación, que les permitan mantener una adecuada relación parental en beneficio de los hijos comunes. Algunos ejemplos los tenemos en la Sentencia de la Audiencia Provincial de Barcelona (Sección 12) n. 146/2014 de 27 de febrero (JUR/2014/85014).

En la Sentencia del Juzgado de 1ª Instancia de Málaga n. 661/2012 de 27 de septiembre (AC/2012/1920) se afirma que los acuerdos alcanzados en mediación tienen un "plus" de obligatoriedad, como una "obligatoriedad reforzada" debido a que se elaboran en un "entorno especialmente apto para que la expresión de la voluntad allí recogida, lo haya sido sin vicio alguno, pues se desarrolla por la intervención técnica del mediador, la voluntariedad de la participación, la igualdad en el desarrollo de los debates que llevan al consenso…".

Sentencias que incluso ponderan la conveniencia de la mediación frente a otras vías como las del Tribunal Supremo (Sala de lo Civil) n. 324/2010 de 20 mayo [RJ\2010\3707]; n. 129/2010 de 5 marzo [RJ\2010\2390]; n. 527/2009 de 2 julio [RJ\2009\6462]; y n. 537/2009 de 3 julio [RJ\2009\5491]

Sentencias como la SAP de Alicante (Sección 14) de 17 de julio de 2015. Sentencia n. 264/2015 (JUR 2015/ 270669) que destaca las ventajas de la mediación, como la mejora de la comunicación y la reducción de los conflictos entre los miembros de la familia, facilitando acuerdos amigables y la preservación de las relaciones personales entre padres e hijos. Sentencia SAP de Barcelona (Sección 18) de 15 de septiembre de 2014. Sentencia n. 379/2014 de 29 de mayo (JUR 2014/227655).

totalmente cercenada y asumida completamente por los poderes públicos. En este sentido son diversos los autores, penalistas, criminólogos y expertos en la materia que han criticado este desapoderamiento de los protagonistas del conflicto en el momento de tomar decisiones sobre las consecuencias del delito que les afectan directamente. Por ejemplo, Nils Christie (1977, 15), en su célebre artículo *"Conflicts as property"*, denuncia y critica la expropiación que se realiza del conflicto, confiscándolo de manos de sus auténticos protagonistas para ceder la gestión de sus consecuencias a las instituciones, profesionales y funcionarios públicos.

Precisamente en la esfera penal ha surgido desde los años setenta del siglo pasado un nuevo paradigma al que el criminólogo y académico Howard Zehr (llamado el padre, y a veces el abuelo, de la justicia restaurativa) bautizó con el nombre de justicia restaurativa en 1985 (Zehr 1991) y que, en la misma línea de Christie, reivindica el papel de las partes y el valor de que el daño causado a la víctima pueda ser reparado por el victimario, frente a la pulsión meramente punitiva.

La justicia restaurativa surge a causa de la insatisfacción de muchos operadores jurídicos, penalistas, criminólogos y académicos respecto al funcionamiento y los resultados del derecho penal. El mismo Zehr considera que el movimiento de la justicia restaurativa surge como un esfuerzo por replantear las necesidades generadas por los delitos. Carencias y necesidades surgidas a raíz de un acto delictivo que el proceso judicial tradicional no estaba atendiendo. Este enfoque en las necesidades de las partes y en los roles que desempeñan ha sido fundamental para este movimiento desde sus inicios (Zehr 2007, 18).

También ha influido el auge de la victimología, que reconoce el papel central de la víctima y el impulso de los MASC como sistemas que reconocen la capacidad de decidir de las personas protagonistas de un conflicto, incluso cuando este conflicto está tipificado como delito. La justicia restaurativa supone un cambio de mirada respecto al derecho penal tradicional. Desde el primer momento el foco no se pone exclusivamente en la pena a imponer al victimario, sino en la posibilidad de que repare a la víctima y en motivar en él un proceso de reflexión y responsabilización, teniendo en cuenta el contexto social en el que se produjo el delito. Las partes protagonistas del conflicto penal son reconocidas y pueden participar en la búsqueda de decisiones y soluciones útiles y consensuadas que neutralicen o compensen los efectos perjudiciales y dañinos del delito. Por tanto, a diferencia de la justicia penal tradicional, la reacción inicial no se focaliza tanto en buscar una respuesta de signo meramente punitivo, sino en facilitar y

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

valorar las posibilidades de reparar a la víctima por parte del propio ofensor, como una oportunidad de carácter voluntario tanto para la víctima, como también para el victimario.

En lugar de "compensar" un daño con otro daño (en la justicia tradicional el perjuicio producido con el delito se intenta equilibrar con la pena establecida y judicialmente impuesta) la filosofía que subyace en la justicia restaurativa es la de subsanar el daño producido con acciones posteriores que sean positivas y sanadoras. En primer lugar, para la víctima que podrá manifestar y satisfacer sus auténticas necesidades, con la finalidad de paliar el mal padecido, y en segundo lugar para el victimario, ofreciéndole la oportunidad de responsabilizarse sobre su propia conducta, resarcir por sí mismo, de forma proactiva, el daño causado y reparar su propia imagen ante la sociedad de la que ambos forman parte (Vall 2022, 50).

La oportunidad que supone la justicia restaurativa, tanto para la víctima como para el victimario incide en la línea de respetar y promover ese derecho humano a poder tomar decisiones, y a que nuestras opiniones y necesidades personales sean tenidas en consideración en escenarios y circunstancias que afectan directamente a nuestra propia vida, especialmente en aquellas situaciones que pueden impactar e incluso alterar, significativamente, nuestro curso vital como es padecer un delito o ser sentenciado a cumplir una pena privativa de libertad.

El mismo Zehr, en su célebre conferencia: *Human rights meets restorative justice* (los derechos humanos se encuentran con la justicia restaurativa) del 20 de diciembre de 2019, pronunciada en un acto organizado por la Carter School for Peace and Conflict Resolution de la Georges Mason University[8], destacó esta relación simbiótica entre la justicia restaurativa y los derechos humanos.

En este punto es conveniente relacionar este derecho personal y humano de los protagonistas a participar también en el diseño de las consecuencias y decisiones que les afectan muy directamente y que deban adoptarse tras la comisión de un delito, con otro derecho humano de reciente consideración: el derecho a equivocarse como derecho humano.

El biólogo y filósofo chileno Humberto Maturana afirma que existen tres derechos humanos universales que no fueron recogidos

[8] Véase: https://www.google.com/search?q=howard+zehr+human+rights&oq=Howard+Zehr+human+&gs_lcrp=EgZjaHJvbWUqBwgBECEYoAEyBggAEEUYOTIHCAEQIRigATIHCAIQIRigATIHCAMQIRigAdIBCTI1ODI4ajBqN6gCALACAA&sourceid=chrome&ie=UTF-8#fpstate=ive&vld=cid:7512ea27,vid:Ccz55SO4Ah4,st:0

por las Naciones Unidas, pero que son tan esenciales como los demás: el derecho a cambiar de opinión, el derecho a irse sin que nadie se ofenda, y el derecho a equivocarse. En palabras de Maturana, estos tres derechos son los que, junto a los demás, hacen posible que un organismo pueda vivir plenamente, construyéndose a sí mismo a lo largo de la vida desde la profunda conexión consigo mismo. Maturana denominó esta teoría como "autopoiesis" combinando dos palabras griegas: "auto" a sí mismo y "poiesis" creación (Maturana y Varela 2003). El derecho a equivocarse es fundamental, porque permite vivir sin el miedo a hacer las cosas "mal", sin la preocupación por no cumplir con las expectativas ajenas. Si, como afirma Maturana, nos podemos equivocar y eso puede ser contemplado como un derecho humano, significa que también debemos tener la oportunidad de poder rectificar y que se nos permita recrear una nueva situación superando el error.

Sin equivocaciones, estaríamos condenados a una eterna repetición de lo mismo. Entiende Maturana que, si percibimos los errores como algo natural en vez de algo irremediable, podemos seguir avanzando y corrigiendo sobre la marcha, manteniendo una actitud inquieta y curiosa ante la vida. La justicia restaurativa a través de sus diversos métodos, como la mediación, los círculos, los círculos de sentencia y las conferencias o encuentros, aporta los instrumentos necesarios, no solo para reconocer y restaurar a la víctima en sus necesidades y derechos conculcados, sino también para asumir este derecho humano a equivocarse y, sobre todo, para reconocer el derecho a poder rectificar, a poder reparar, a autorepararse, personal y socialmente, y a tener la oportunidad de avanzar y recrear un nuevo curso vital en el sentido autopoietico de Maturana.

Con ello se puede contribuir a superar procesos criminógenos y la asunción de etiquetas (proceso de etiquetamiento social como delincuente o *labelling approach*[9]). La creación de estereotipos y la estigmatización social vinculada al paradigma meramente punitivo no aportan reflexión ni cambios en la persona victimaria, sino que despiertan reactividad, adaptación al etiquetado e impulsan negativamente la deriva hacia la reincidencia.

[9] La teoría del etiquetado se enmarca dentro de la sociología de la desviación y según Howard Becker (2009): "la desviación no es una cualidad de la acción cometida, sino la consecuencia de la aplicación por parte de otros de reglas y sanciones". El desviado es alguien al que la etiqueta le ha sido puesta, el comportamiento desviado es el comportamiento etiquetado de una manera concreta por otros. Véase:

https://www.unir.net/revista/derecho/teoria-de-etiquetamiento/

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

Respecto a la víctima, su participación en programas de justicia restaurativa supone la posibilidad de ser reparada material y/o moralmente, no únicamente en base a lo que la ley pueda considerar adecuado como compensación punitiva genérica, sino en función de satisfacer las necesidades concretas surgidas a raíz de la comisión del delito que la propia víctima tiene la posibilidad de identificar y reclamar directamente al victimario. Al mismo tiempo la víctima recupera su autoestima, su dignidad y refuerza su sentido de autonomía al poder tomar decisiones también sobre las consecuencias de los hechos delictivos vividos, reforzando así su capacidad de superación de las secuelas del delito y potenciando su resiliencia y su proceso de desvictimización. Por ello también respecto a la víctima, para ayudarla en su recuperación y en el camino de superación del trauma vivido a raíz del delito, la posibilidad de optar por una respuesta restaurativa debería entenderse como un derecho humano al que poder acceder en igualdad de condiciones por parte de todas las personas que han sido víctimas de un delito. Sin embargo y pese a sus señaladas ventajas personales y sociales, tal como indica Tamarit (2020, 43), las prácticas restaurativas desempeñan actualmente un papel limitado y en la mayoría de los países son solo una realidad marginal entre las formas de respuesta a la delincuencia.

Frente a este rol secundario de las respuestas restaurativas, la administración pública debería velar para que este acceso al derecho fundamental de la víctima, no solo a ser reparada, sino también a participar en la toma de decisiones sobre cómo ha de ser esa reparación, sea aplicable en la realidad de la práctica a todos los supuestos en que la víctima libremente lo solicite y opte por ello, sin que deba producirse ningún tipo de discriminación o impedimento en su accesibilidad por razón de su edad, condición o lugar en el que resida.

Este derecho de la víctima a ser reparada participando directamente en un programa restaurativo queda recogido en nuestro país en el artículo 15 de la Ley 4/2015 del Estatuto de la Víctima del Delito, que contempla la existencia de los servicios de justicia restaurativa y el acceso a estos servicios como un derecho de las víctimas condicionado a la concurrencia de determinados requisitos.

El art. 15 de la Ley del Estatuto de la Víctima del Delito establece lo siguiente:

> 1. Las víctimas podrán acceder a servicios de justicia restaurativa, en los términos que reglamentariamente se determinen, con la finalidad de obtener una adecuada reparación material y moral de los

perjuicios derivados del delito, cuando se cumplan los siguientes requisitos:

a) el infractor haya reconocido los hechos esenciales de los que deriva su responsabilidad;

b) la víctima haya prestado su consentimiento, después de haber recibido información exhaustiva e imparcial sobre su contenido, sus posibles resultados y los procedimientos existentes para hacer efectivo su cumplimiento;

c) el infractor haya prestado su consentimiento;

d) el procedimiento de mediación no entrañe un riesgo para la seguridad de la víctima, ni exista el peligro de que su desarrollo pueda causar nuevos perjuicios materiales o morales para la víctima;

e) no esté prohibida por la ley para el delito cometido.

2. Los debates desarrollados dentro del procedimiento de mediación serán confidenciales y no podrán ser difundidos sin el consentimiento de ambas partes. Los mediadores y otros profesionales que participen en el procedimiento de mediación estarán sujetos a secreto profesional con relación a los hechos y manifestaciones de que hubieran tenido conocimiento en el ejercicio de su función.

3. La víctima y el infractor podrán revocar su consentimiento para participar en el procedimiento de mediación en cualquier momento.

Las funciones básicas de estos servicios se regulan en el art. 19 del Real Decreto 1109/2015, por el que se desarrolla la Ley 4/2015 del Estatuto de la Víctima del Delito. Dicho artículo 19 del Real Decreto regula, entre las funciones de las Oficinas de Asistencia a las Víctimas, la de proporcionar a las víctimas "información sobre alternativas de resolución de conflictos con aplicación, en su caso, de la mediación y de otras medidas de justicia restaurativa".

La Ley 4/2015 del Estatuto de la Víctima del Delito y su Reglamento de desarrollo suponen una traslación a nuestro derecho interno de la Directiva 2012/29/UE del Parlamento Europeo y del Consejo de 25 de octubre de 2012, por la que se establecen normas mínimas sobre los derechos, el apoyo y la protección de las víctimas de delitos y que sustituye la Decisión Marco 2001/22/JAI del Consejo. Se trata, por tanto, de una disposición europea que afecta y vincula normativamente a todos los Estados miembros de la Unión Europea.

La administración pública debe considerar y ser consciente de las nuevas inquietudes sociales y articular los mecanismos necesarios para que estos derechos humanos de víctimas y victimarios, protagonistas en el ámbito penal, puedan ser reconocidos y hacerse efectivos más allá de la pulsión únicamente punitiva. De hecho, nuestra propia Constitución en su art. 25.2 establece que, incluso, las penas privativas

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

de libertad y las medidas de seguridad estarán orientadas hacia la reeducación y reinserción social, marcando así un objetivo principal no de carácter punitivo sino recuperador y resocializador.

Estos mecanismos de reconocimiento del valor de las personas y de su capacidad para decidir por ellas mismas, que como hemos visto, pueden operar en distintos ámbitos de las relaciones interpersonales (familia, civil, mercantil, penal…) han sido denominados con el acrónimo MASC[10]. La letra A puede entenderse como A de sistemas apropiados o adecuados, aunque inicialmente se interpretaba como alternativos al sistema judicial, en la línea del originario acrónimo anglosajón ADR (*Alternative Dispute Resolution*). Actualmente se considera que más que alternativos pueden considerarse adecuados, en un doble sentido: primero porque no se trata de una alternativa fuera del sistema judicial sino de un complemento dentro del sistema (el propio juez en un caso judicializado puede derivar las partes a un profesional o a una institución que aplique uno de estos medios) y, segundo, porque es la decisión judicial la que puede ser alternativa en muchos casos, priorizando que, en primer lugar, sean las propias partes las que traten de alcanzar acuerdos y soluciones y, solo de forma subsidiaria, el juez tenga que decidir a falta de consenso entre las partes. Como hemos visto, bajo la apelación de MASC se engloban distintos sistemas como la mediación, la conciliación o la negociación que son métodos autocompositivos inspirados en una filosofía común que, al igual que en la justicia restaurativa reconocen un mayor protagonismo a las personas que viven una situación de conflicto. Los propios protagonistas del conflicto adquieren también un papel central en el proceso de toma de decisiones acerca de cómo superar estas situaciones por ellas mismas, sin que las respuestas formales, que muchas veces no son auténticas soluciones, sino una aplicación genérica del derecho, les vengan impuestas por un tercero.

Kant decía que es difícil salir de la minoría de edad, porque estamos a gusto en ese estado por comodidad y por cobardía. Puede parecer más cómodo que alguien piense y decida por mí, ya que eso nos evita responsabilidades. Para Kant la minoría de edad significa la incapacidad para servirnos de nuestro propio entendimiento. En este sentido, Kant (2012) popularizó la expresión latina *Sapere aude!* (atrévete a pensar) que había tomado de Horacio[11]. La mayoría de

[10] MASC es la denominación que utiliza también el comentado Proyecto de Ley de Eficiencia Procesal del Servicio Público de Justicia para referirse a estos métodos.

[11] La expresión original *Sapere aude*! aparece en una epístola de Horacio a su amigo Lollius Maximus del año 20 a.de C. El filósofo alemán Immanuel Kant (1784)

edad supone la capacidad de obrar, la capacidad de ejercer por uno mismo los derechos y obligaciones de los que es titular. La capacidad de servirse de su propio entendimiento, tener el valor de utilizar tus habilidades para pensar y por supuesto, para tomar tus propias decisiones. En este punto de nuestra civilización, en los países que pueden considerarse como estructurados bajo la forma de un estado social y democrático de derecho, la inmensa mayoría de ciudadanos podemos considerar que hemos alcanzado la mayoría de edad y que estamos preparados para pensar, ser conscientes de nuestra realidad y tomar nuestras propias decisiones, especialmente en aquellos asuntos que afectan directamente a nuestra vida y que tienen trascendencia en nuestro presente, en nuestro futuro y en las relaciones con nuestro entorno más inmediato.

Si vivimos por tanto en sociedades adultas, nuestro marco jurídico ha de reconocer la capacidad de las personas para poder tomar sus propias decisiones y el derecho ha de acompañar y facilitar esa toma de decisiones autónomas, a través del reconocimiento y la articulación de los mecanismos apropiados para garantizar la posibilidad de que todos los ciudadanos puedan utilizar, de forma voluntaria, estos métodos como un derecho humano universalmente accesible, sin que quepa ningún tipo de discriminación, más allá de las condiciones que puedan establecerse por ley para garantizar la efectividad y la correcta aplicación de tal derecho dentro del marco legal vigente, velando por el cumplimiento y la armonización con los demás derechos personales y ciudadanos reconocidos por nuestro ordenamiento.

Las distintas administraciones públicas deben facilitar esta opción por igual a todas las personas, por tratarse de un derecho que no debe discriminar a ningún ciudadano, ya sea en función de sus rasgos personales, de su capacidad económica o de su residencia habitual. En este sentido se trata de una forma de justicia en la que las partes adquieren un mayor protagonismo. Las mismas disposiciones europeas y españolas lo catalogan como una forma de justicia de calidad[12], ya que implica a los propios protagonistas y les reconoce un papel central en el proceso de toma de decisiones que, posiblemente sean más satisfactorias, eficientes, sin duda alguna consensuadas y aceptables por parte de todos.

---

divulgó esta locución en su ensayo ¿Qué es la Ilustración?, en el sentido expresado de atrévete a hacer uso de tu propio entendimiento.

[12] El preámbulo de la Ley 5/2012 de Mediación en asuntos civiles y mercantiles, remarca que una de las funciones del Estado es "la implantación de una justicia de calidad" en referencia a la mediación.

Por ello, las soluciones que se alcanzan a través de estos MASC suelen ser más sólidas en el sentido de obtener un mayor grado de cumplimiento y perdurabilidad temporal. Esta mayor efectividad se debe a que las partes se sienten directamente concernidas, por los compromisos acordados, ya que ellas mismas han participado directamente en su elaboración, consensuando decisiones que han sido fruto de sus propios planteamientos, ponderando intereses y necesidades comunes e individuales en base a su conocimiento de la concreta realidad que ellos viven en primera persona.

3. **Los recursos de la Inteligencia Artificial como facilitadores de los MASC**

Si nos acercamos al significado de la Inteligencia Artificial (IA), ésta puede ubicarse en el campo de la informática que se enfoca en crear sistemas que puedan realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, el razonamiento y la percepción.

El 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 IA y por el que se modifican los Reglamentos (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 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) define en su art. 3 el sistema de IA 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 o virtuales".

La IA supone, por tanto, la simulación de inteligencia humana que crea algoritmos y sistemas informáticos capaces de ejecutar tareas simples y complejas que realizan las personas. Se basa en la idea de que una máquina puede programarse para imitar la forma en que un ser humano piensa y actúa. La IA no es un ideal de futuro, sino que su penetración en nuestra vida cotidiana actual es evidente y útil en múltiples campos y aspectos (gestión del tráfico, vehículos autónomos, asistentes virtuales, gestión administrativa, automatización de procesos industriales, investigación…) e incide en nuevas formas de trabajar y de gestionar los problemas y la vida de las personas y de las empresas.

Quizá una de las preguntas a formular es si algún día la IA pueda llegar a superar la inteligencia humana y actuar de forma totalmente autónoma e independiente del ser humano y los riesgos que eso podría implicar para la humanidad ¿Puede llegar a desarrollar una voluntad propia? ¿Podría adoptar decisiones independientes de cualquier programación previa humana? ¿La IA puede abordar y asumir cuestiones de tipo ético?

En este sentido el mencionado Reglamento (UE) 2024/1689 establece en su art. 14 la importancia de la supervisión humana de los mecanismos de la IA y concretamente en su apartado segundo nos dice que "el objetivo de la supervisión humana será prevenir o reducir al mínimo los riesgos para la salud, la seguridad o los derechos fundamentales…". Por otra parte, este mismo Reglamento en su considerando 61 clasifica como de "alto riesgo" aquellos sistemas de IA "destinados a la administración de justicia y los procesos democráticos, dado que pueden tener efectos potencialmente importantes para la democracia, el Estado de Derecho, las libertades individuales y el derecho a la tutela judicial efectiva…". Cabe por tanto deducir que el propio Reglamento es consciente de que una inadecuada utilización de los mecanismos de la IA podría suponer un riesgo cierto para los pilares y los valores fundamentales de nuestras sociedades, basadas en criterios democráticos y en el respeto hacia los derechos fundamentales. De ahí la importancia de una supervisión humana que además sea rigurosa en evitar sesgos discriminatorios y que alimente los recursos de la IA con datos e informaciones veraces y contrastadas, basada en la ética, los valores democráticos y los derechos humanos.

Como recuerda Gema Varona (2020), la utilización de algoritmos se realiza en base a datos masivos producidos y recopilados por seres humanos que, por tanto, pueden tener sesgos y prejuicios. Por su parte, Jorge Sánchez López comenta que la IA no es una simple creación tecnológica, sino que refleja valores y aspiraciones que proyectamos como sociedad y se adapta a los objetivos que se le asignen. Si la IA se programa para la dominación y el engaño se convierte en un reflejo de las peores intenciones, si por el contrario se orienta hacia la cooperación y la solidaridad puede proporcionar una herramienta muy útil para el progreso humano (Sánchez López 2024)[13]

En todo caso la IA puede suministrar una amplia gama de instrumentos aplicables en distintos campos y ámbitos que estén

[13] Véase: https://www.linkedin.com/pulse/la-ia-un-espejo-de-nuestra-alma-jorge-s%C3%A1nchez-l%C3%B3pez-usw9f/

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

dispuestos al servicio de las personas para contribuir a una mayor calidad de vida y para facilitar nuestras actividades tanto en el ámbito de la investigación académica, como en los aspectos personales, relacionales, sociales, laborales, sanitarios o profesionales.

A pesar de ello, y de momento, son muchas las acciones y reacciones que todavía pueden considerarse exclusivas del ser humano, como las emociones, los sentimientos, los valores, las ideas, los conceptos éticos, religiosos o filosóficos de la vida, las percepciones, etc.

Los conflictos y discrepancias entre las personas y los grupos humanos se construyen a partir de un amplio espectro de matices y bajo la influencia de múltiples emociones, sentimientos, percepciones, puntos de vista, experiencias y vivencias previas que concurren en el origen y desarrollo de los conflictos interpersonales. Esta complejidad hace que difícilmente estas discrepancias y conflictos que, apelan no solo a la capacidad de razonar, sino que, en buena medida, están tejidos con las emociones y sentimientos íntimos y propios del ser humano, puedan gestionarse y encontrar respuestas satisfactorias directamente a través de la aplicación de un sistema de IA totalmente mecánico, independiente y autónomo del acompañamiento humano.

Martín Diz (2020) señala que la utilización de la IA en la solución extrajudicial de conflictos puede desempeñar dos grandes funciones: la función asistencial o la función decisoria. Desde estas líneas se coincide totalmente en distinguir la posibilidad de esta doble funcionalidad de la IA respecto a la gestión extrajudicial de los conflictos. Pero, así como la función asistencial es evidente, necesaria, totalmente útil, incluso a veces imprescindible en la práctica, especialmente desde la pandemia de 2020, adjudicar la función decisoria a la IA, en este momento (no podemos avanzar lo que pasará en el futuro), aparece como una posibilidad más compleja, polémica y quizá utópica. Como se pregunta el mismo autor "¿Arbitrará o mediará un litigio una IA? ¿Alcanzarán robots, avatares y otros agentes relacionados o softwares el nivel de confianza suficiente en los litigantes que tienen los árbitros y mediadores humanos?"

En línea con esta segunda posibilidad decisoria, Gema Varona (2020) relata que, en países como China o Estonia, entre otros, existen plataformas en los tribunales en las que, sin intervención humana, las partes cargan los datos de los conflictos a resolver y, a través de mecanismos de la IA, se busca jurisprudencia, se contrastan pruebas y se emite una resolución. De todas formas, estas resoluciones posiblemente sean estandarizadas y difícilmente tengan en consideración la multiplicidad de factores y la diversidad de matices emocionales y psicológicos que tanto inciden en la génesis, evolución y

superación de cada conflicto específico y concreto. Será imprescindible la realización de estudios e investigaciones que analicen estas resoluciones elaboradas directamente a través de mecanismos de la IA y, sobre todo, será fundamental promover una investigación y un seguimiento posterior de los niveles de satisfacción o insatisfacción generados por su aplicación práctica y sus resultados a corto y a largo plazo.

Parece preocupante y desconocido el riesgo que podría provocar la incorporación de esta posible función decisoria de la IA en el ámbito de la Justicia. Por ello el mencionado Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo tiene el objetivo de crear un marco normativo que preserve especialmente de aquellas situaciones que puedan afectar a los derechos fundamentales.

En todo caso, la IA no queda al margen, sino que, al contrario, juega ya un importante papel en el abordaje consensual y eficaz de los conflictos como medio facilitador de esa búsqueda de soluciones acordadas en las que se implican las propias partes con la ayuda de un tercero y la colaboración de la IA que, en numerosas ocasiones, crea el marco idóneo y la vía de comunicación efectiva y propicia para hacerlo posible.

Los distintos recursos de la IA pueden facilitar que la opción de la puesta en práctica de un método MASC (mediación, conciliación, negociación, círculos restaurativos, etc.) como una forma de resolver controversias y aplicar justicia, sea posible, especialmente cuando concurre una determinada coyuntura que dificulta el encuentro presencial. Circunstancias como la distancia física entre las partes, la dificultad de desplazarse, el deseo de no coincidir físicamente, una orden de alejamiento que no comporte la prohibición de comunicarse (ya que la puesta en práctica de estos métodos se basa justamente en fomentar una comunicación adecuada y hacer posible el diálogo entre las partes). En estas situaciones y otras, la IA puede convertirse en un "colaborador necesario" como una vía facilitadora para que la opción de los MASC sea posible y accesible en la práctica, para todas las personas más allá de sus circunstancias y dificultades concretas. El acrónimo MASC en el mundo anglosajón es más conocido como ADR (Adequate Dispute Resolution) que unido a la IA empieza a ser conocido con la nueva nomenclatura de AIDR, enlazando así ambos conceptos.

Estos mismos métodos de gestión adecuada de conflictos, cuando se llevan a la práctica a través de medios de comunicación tecnológicos no presenciales, propios de la IA, se les denomina Métodos ODR, por sus iniciales en inglés (*On-line Dispute Resolution*). La propia web de la

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

Unión Europea facilita la resolución de litigios en línea (métodos ODR) para gestionar y resolver las reclamaciones en materia de consumo, que pueden gestionarse directamente por el mismo usuario afectado[14].

Además de distintas normas internacionales sobre ODR, como el Reglamento (UE) n. 524/2013 del Parlamento Europeo y del Consejo de 21 de mayo de 2013, sobre resolución de litigios en línea en materia de consumo, debe mencionarse la Directiva 2008/52/CE de la Unión Europea que contiene diversas disposiciones que reflejan la voluntad de potenciar estos métodos en el sentido expresado en este texto.

Así el considerando 5º de dicha Directiva establece como objetivo prioritario asegurar un mejor acceso a la justicia como parte de la política de la UE que debe incluir "el acceso a métodos tanto judiciales como extrajudiciales de resolución de litigios… en particular en lo referente a la disponibilidad de servicios de mediación".

Por su parte el considerando seis establece que "la mediación puede dar una solución extrajudicial económica y rápida a conflictos en asuntos civiles y mercantiles, mediante procedimientos adaptados a las necesidades de las partes. Es más probable que los acuerdos resultantes de la mediación se cumplan voluntariamente y también que preserven una relación amistosa y viable entre las partes".

Teniendo en cuenta estas ventajas que describe en sus considerandos, el artículo primero reconoce como objetivo de la Directiva el de facilitar el acceso de todas las personas a las modalidades alternativas de solución de conflictos y el fomento de la resolución amistosa de litigios, asegurando una relación equilibrada entre la mediación y el proceso judicial. Destaca también, claramente, la ineludible adaptación de los procedimientos a las necesidades de las partes. Al no concretar cuales son estas necesidades puede interpretarse de forma amplia que se trata de atender, en la medida de lo posible, lo que requieren las personas en función de las características particulares del caso concreto.

En ocasiones pueden existir circunstancias geográficas, complicaciones comunicativas, imposibilidad de los encuentros presenciales o impedimentos de otro tipo que dificulten este acceso a los métodos MASC. Estas limitaciones deben y pueden ser superadas, tal como se expresa en el art. 9 de la misma Directiva 2008/52 UE al señalar que "los Estados miembros fomentarán, por los medios que

[14] Véase: https://ec.europa.eu/consumers/odr/main/index.cfm?event=main.home2.show&lng=ES

consideren oportunos, el acceso del público en general, en particular vía internet, a la información sobre la forma de ponerse en contacto con mediadores y organismos que presten servicios de mediación".

Lo dispuesto en este artículo 9 es coherente con el contenido del considerando noveno que establece que "la Directiva no debe impedir en modo alguno la utilización de las nuevas tecnologías de comunicaciones en los procedimientos de mediación". Con lo cual los procedimientos y recursos de la IA y de las nuevas tecnologías no solo se reconocen como mecanismos útiles para facilitar la información sobre la existencia y el acceso a la mediación y a los MASC en general, sino que además pueden ser utilizados para poner en práctica el mismo proceso de mediación o bien otros medios adecuados de solución de controversias.

Precisamente la pandemia del COVID-19 supuso un punto de inflexión en la utilización de diferentes canales digitales que permitieran superar la imposibilidad o dificultad para el encuentro físico debido a las limitaciones impuestas a la movilidad de las personas, especialmente en el período más duro del aislamiento. Durante la pandemia muchas personas mediadoras, utilizando distintos programas, seguimos realizando procesos de mediación, ya fuese para finalizar aquellas mediaciones que ya estaban en marcha o bien dando inicio a otras, aunque de forma no presencial. Desde distintas administraciones públicas locales y autonómicas se optó también por facilitar la gestión telemática, como el Centro de Mediación de la Generalitat de Cataluña, que auspició un Programa de reuniones telemáticas durante la pandemia para gestionar las mediaciones familiares y ciudadanas[15] o el Servicio de mediación penal del Gobierno de Navarra, gestionado por la Asociación ANAME que continuó con su actividad de forma telemática, o el Servicio de Mediación Comunitaria del Consejo Comarcal del Alt Penedès, gestionado por Logos Media, entre otros muchos, que siguieron desarrollando su labor mediadora y facilitadora a través de distintos recursos virtuales. A partir del COVID, muchos de estos procesos siguen realizándose, opcionalmente, de forma telemática, si la persona mediadora y las partes así lo acuerdan por ser más fácil o preferible para los protagonistas del caso concreto.

Pero no únicamente es posible realizar mediaciones de forma digital, sino que también pueden activarse otro tipo de metodologías MASC. Gema Varona (2020) pone como ejemplo, el National Conflict

[15] Resolución JUS 848/2020 de 1 de abril del Departamento de Justicia de la Generalitat de Cataluña, que acuerda el seguimiento de los procedimientos de mediación del Centro de Mediación.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

Resolution Center de EE.UU. que en 2020 llevó a cabo círculos comunitarios virtuales utilizando *Zoom*, para reunir a personas que atravesaban distintas situaciones conflictivas; y refiere Varona que estos círculos virtuales consiguieron incentivar el diálogo y la creación de empatía entre los participantes.

La aplicación de los MASC en formato digital o telemático presenta ventajas y también inconvenientes en comparación con su puesta en práctica presencial. Como ventajas debe mencionarse la facilidad para propiciar un diálogo que en ocasiones no sería factible de forma presencial, ya sea por la distancia física, por falta de tiempo para desplazarse, porque la presencialidad incomoda a alguna de las partes, por sentirse las personas más seguras o cómodas al poder llevar a cabo el proceso desde su propio espacio o zona de confort, o bien para tener a mano documentos, fotos u otros elementos que quieran mostrarse desde la pantalla, etc. En otras ocasiones, a través del formato on-line, las partes consiguen expresar o compartir cosas o vivencias que presencialmente no se atreverían.

Como inconvenientes debe mencionarse que la presencialidad tiene un plus comunicativo en el sentido de que todo nuestro cuerpo, nuestros gestos, expresiones y micro expresiones transmiten emociones, sensaciones, percepciones… Por otra parte, las dificultades tecnológicas o de conexión de las personas o del entorno pueden ser también un impedimento en algunas situaciones. En todo caso, si no se da una situación imperativa como durante la pandemia, aplicar los MASC en formato digital debería ser una opción abierta a todos los ciudadanos, pero nunca una obligación, ni un único camino para acceder a los MASC.

Este artículo parte de concebir el acceso a los métodos MASC como un derecho de todos los ciudadanos y de todas las personas en general, un derecho que permite optar por otras vías diferentes, pacíficas y complementarias a la vía judicial. No se trata de salir del sistema, sino todo lo contrario, dentro del mismo marco jurídico, deben existir los recursos y mecanismos necesarios para ofrecer respuestas no contenciosas, más eficientes, útiles y adecuadas a cada conflicto y situación. Si situamos el objetivo en conseguir que esta opción pueda ser efectiva y real para todas las personas sin ningún tipo de limitación o discriminación, la IA y sus múltiples recursos juegan un papel fundamental, como vías facilitadoras, para hacer realidad este objetivo.

Conclusiones y recomendaciones

I. Los MASC suponen una nueva forma de gestionar los conflictos y las discrepancias entre las personas, ya sean físicas o jurídicas. Estos métodos reconocen el papel central de las partes de un conflicto, que se convierten también en protagonistas en la búsqueda de soluciones consensuadas que sean vividas como justas, apropiadas, realistas y factibles para todos. Por ello numerosos expertos, como Lauroba y Ortuño (2018), entienden que estos métodos suponen una nueva forma de hacer Justicia y que las propias personas implicadas son las que mejor conocen su realidad y, por tanto, son también las que pueden construir las soluciones más justas, satisfactorias y ajustadas a cada caso y situación. Todo ello, como señala Ortuño (2018) sin que la no participación de un Juez en la toma de decisiones suponga, en absoluto, una merma en la calidad de la respuesta a la controversia o que disminuya en las partes la vivencia de justicia de la solución acordada que se adopte finalmente.

II. La voluntariedad y la participación directa de los implicados es fundamental para gestionar los conflictos y discrepancias a través de alguno de estos medios MASC, y también la colaboración de una tercera persona que facilite el diálogo y el consenso sin imponer, en ningún caso, una solución, que no sea la que encuentren conjuntamente las propias partes. Por ello el factor humano es fundamental, ya que en la generación de la mayoría de los conflictos y discrepancias suele tener una elevada incidencia y protagonismo la faceta emocional, superando muchas veces la capacidad de gestionar basada puramente en lo racional. La comprensión de estos factores vivenciales que envuelven el conflicto y condicionan su desarrollo y efectos requiere, en principio, del acompañamiento de un profesional capaz de identificar y canalizar dichas emociones y sentimientos, que escapan muchas veces del análisis puramente racional.

III. Esta mezcla de matices y emociones humanas, a veces inconscientes por parte de quien las siente o de quien las provoca, hace necesaria la intervención de un tercero que genere confianza y despierte la posibilidad de motivar una comprensión mutua y la generación de nuevos escenarios en los cuales, más allá de la confrontación, tenga cabida la colaboración, la creatividad y la búsqueda de soluciones compartidas. Por tanto, difícilmente ningún mecanismo de la IA, en este momento, puede asumir la labor del tercero profesional que aplica una fórmula MASC y que no parte de una "programación concreta", ni del funcionamiento de determinados algoritmos, sino que acompaña a las personas centrándose en la

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

comprensión de los múltiples, diversos y complejos procesos mentales y emocionales concurrentes en cada caso, que se ven altamente potenciados y polarizados en situaciones de conflicto.

IV. Si bien es improbable que actualmente la IA pueda sustituir al profesional que gestiona una situación conflictiva compleja a través de un método MASC, o que pueda asumir funciones decisorias autónomas, sí que puede llegar a tener un papel muy destacado como canal facilitador de la aplicación de estos métodos y ser clave para posibilitar la participación activa de las personas protagonistas junto al tercero profesional. A partir de la pandemia a nivel mundial del COVID-19, el desarrollo de la comunicación a través de distintos sistemas virtuales se ha incrementado de forma exponencial, tanto para realizar actividades y comunicaciones de tipo familiar como de ámbito docente, profesional, comercial, empresarial y también en la aplicación virtual de la mediación y de otras fórmulas MASC.

V. Ya actualmente muchos de los procesos de gestión colaborativa de conflictos (MASC) que se llevan a cabo entre personas o entre distintas entidades y empresas se implementan con la ayuda imprescindible de sistemas de comunicación que forman parte de la IA, como los denominados ODR *(On line Dispute Resolution)*, con una función básicamente asistencial no decisoria. Sería recomendable una formación adecuada de los profesionales de los MASC para poner en práctica estos procesos, no solo presencialmente, sino también en formato virtual para incorporar nuevas habilidades adaptadas a las peculiaridades del medio. En necesario avanzar en la mejora de la forma de comunicarnos virtualmente para que la expresión de tantos matices gestuales, de entonación o expresión o incluso de los silencios puedan transmitirse, comprenderse y valorarse de forma similar a como podemos captarlos en modo presencial.

VI. La IA puede ser el instrumento adecuado para conseguir hacer efectivo el derecho a que todos los ciudadanos tengan un acceso real a la opción de aplicar uno de estos medios de gestión pacífica, colaborativa y eficiente de conflictos, sin ningún tipo de discriminación, ya sea geográfica o material. Si bien la red judicial está extendida por todos los países occidentales y llega generalmente a todos los rincones del Estado, la posibilidad de acudir a un método MASC todavía no queda asegurada en la práctica de la misma forma. Este diferente nivel de accesibilidad de los ciudadanos entre el sistema de justicia tradicional y un sistema de MASC puede venir motivado por falta de una buena red de profesionales, por falta de información, por falta de recursos estructurales o económicos o por otros motivos. Estas dificultades podrían superarse a través de la IA y de sus múltiples recursos comunicativos.

VII. Por todos estos motivos, la implementación de los sistemas de IA puede convertirse en una herramienta muy eficaz para asegurar que todos los ciudadanos tengan acceso a la información y a la posibilidad de optar y aplicar alguno de los métodos MASC, como un derecho esencial a la utilización efectiva de vías diferentes a la contenciosa que faciliten soluciones pacíficas, eficientes y satisfactorias para todos. Si la IA consigue expandir, facilitar y socializar la opción de acudir a alguno de estos medios MASC, estaremos contribuyendo a construir una sociedad más pacífica, madura y responsable. Se evitará así que los ciudadanos, en muchos casos, tengan que acudir necesariamente al aparato judicial contencioso para dejar que un tercero decida sobre cómo solucionar sus discrepancias. Esta asunción de capacidad decisoria y resolutoria por parte de los mismos protagonistas es especialmente deseable en aquellas controversias que tienen una destacada carga conflictual, emocional y relacional con escasa trascendencia jurídica. Con ello, las personas tendrán la opción de construir soluciones consensuadas, útiles y más ajustadas a sus necesidades reales. Indirectamente, el sistema judicial también puede ganar en calidad, tiempo y eficacia para dedicarse a cuestiones de cariz netamente jurídico y de difícil resolución por las propias partes.

Referencias

Becker, Howar. 2009. *Outsiders. Hacia una Sociología de las Desviación* [1963]. Buenos Aires: Siglo XXI Editores.

Christie, Nils. 1977. «Conflicts as property». *The British Journal of Criminology* 17 (1): 1-15.

Kant, Immanuel. 2012. *Contestación a la pregunta: ¿Qué es la Ilustración?* Barcelona: Taurus

Lauroba, M. Elena y Pascual Ortuño, coord. 2014. *Mediación es justicia. El impacto de la Ley 5/2012, de mediación civil y mercantil*, Barcelona: Huygens.

Martin Diz, Fernando. 2020. «Inteligencia Artificial y ADR evolución en el arbitraje y la mediación», *La Ley. Mediación y arbitraje:* 2

Maturana, Humberto y Francisco Varela. 2003. *De máquinas y seres vivos. Autopoiesis: la organización de lo vivido*, Buenos Aires: Lumen.

Ortuño, Pascual. 2018. *Justicia sin jueces: Métodos alternativos a la justicia tradicional.* Barcelona: Ariel.

Sánchez López, Jorge. 2024. *La IA: un espejo de nuestra alma*. Acceso el 1 de septiembre de 2024. https://www.linkedin.com/pulse/la-ia-un-espejo-de-nuestra-alma-jorge-s%C3%A1nchez-l%C3%B3pez-usw9f/

Tamarit, Josep M. 2020. «El lenguaje y la realidad de la justicia restaurativa», *Revista de Victimología* 10: 43-70.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 259-285
https://doi.org/10.18543/djhr.3196 • http://djhr.revistas.deusto.es/

Vall, Anna. 2022. *Justicia restaurativa. Estado de la cuestión y propuestas de lege ferenda*. Valencia: Tirant Lo Blanch.

Varona, Gemma. 2020. «Justicia restaurativa digital, conectividad y resonancia en tiempos del COVIV-19». *Revista de Victimología* 10: 9-42

Zehr, Howard. 2007. *El pequeño libro de la justicia restaurativa*. New York: Good Books.

Zehr, Howard. 1991. *Changing lenses: New focus for crime and justice*. Scottdale: Herald Press.

La inocencia de la responsabilidad social corporativa para proteger los derechos humanos ante la inteligencia artificial

The candor of corporate social responsibility to safeguard human rights from artificial intelligence

Raúl López González
Universidad Autónoma de Madrid. España
raul.lopez.gonzalez@icloud.com
ORCiD: https://orcid.org/0009-0001-6341-6834

https://doi.org/10.18543/djhr.3197

Fecha de recepción: 30.05.2024
Fecha de aceptación: 07.09.2024
Fecha de publicación en línea: diciembre de 2024

Cómo citar / Citation: López González, Raúl. 2024. «La inocencia de la responsabilidad social corporativa para proteger los derechos humanos ante la inteligencia artificial». *Deusto Journal of Human Rights*, n. 14: 287-312. https://doi.org/10.18543/djhr.3197

Sumario: 1. Ética y tecnología. 2. Ética y capitalismo. 3. Derecho a la privacidad. 4. Derecho al trabajo. Conclusiones. Referencias bibliográficas.

Resumen: La ética, la filosofía y la ciencia han permitido a la humanidad resolver los desafíos más complejos a lo largo de su historia. La dignidad humana es el fin ulterior de una civilización, tanto el Renacimiento como la Ilustración sintetizan lo mejor de la ciencia y el pensamiento. Por la ciencia y la dignidad humana llegaron la genética, la digitalización, la neurociencia y la computación cognitiva, que trasladan hoy la civilización humana hacia la cuarta revolución industrial. En la dignidad humana se hace imperativo reconocer el legado auténtico de Adam Smith. Sus mayores críticos no han leído su teoría del sentimiento moral, y sus mayores defensores desacreditan que Smith fuera ante todo humanista, luego economista. Los directivos más avezados abrazarán un capitalismo comprometido con la dignidad humana, sin complejos. Este ensayo enlaza capitalismo y tecnología, para enaltecer la privacidad y el trabajo digno.

Palabras clave: Filosofía, ética, computación cognitiva, humanismo, inteligencia artificial, derechos humanos, negocios, capitalismo.

Abstract: Ethics, philosophy and science have allowed humanity to solve the most complex challenges throughout its history. Human dignity is the

ultimate goal of a civilization, both the Renaissance and the Enlightenment synthesize the best of science and thought. For science and human dignity came nuclear energy, genetics, digitalization, neuroscience and cognitive computing, which today move human civilization towards the fourth industrial revolution. In human dignity, it is imperative to recognize the authentic legacy of Adam Smith. His greatest critics have not read his theory of moral sentiment, and his greatest defenders discredit that Smith was first and foremost a humanist, then an economist. The most seasoned managers will embrace a capitalism committed to human dignity, without complexes. This essay links capitalism and technology, to extol privacy and decent work.

Keywords: Philosophy, ethics, cognitive computing, humanism, artificial intelligence, human rights, business, capitalism.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

1. **Ética y tecnología**

El filósofo clásico Aristóteles sintetizó la ética en prudencia y convivencia (Aristóteles n.d.). Él mismo apadrinó el Liceo, la primera universidad conocida en la que el conocimiento de la biología, geometría y astronomía se prestó al servicio del ser humano, otra forma de ilustración helénica. En su tiempo, la política y la ética eran unívocas, una simbiosis censurada por corrientes teocráticas durante los veinte siglos posteriores a su vida; teocracias que incubaron la opresión y miseria del feudalismo, el absolutismo, las cruzadas religiosas, y los totalitarismos recientes, que aún generan hambre y guerras.

La filosofía será a la razón lo que la ciencia al desconocimiento. Voltaire, filósofo y abogado, es el emblema de la Ilustración, una corriente de pensamiento que surgió a mediados del siglo XVIII. Gracias al pensamiento de Voltaire, la ciencia aceleró su desarrollo en servicio a la humanidad. "Cuando la filosofía ha empezado a ilustrar un poco a los hombres, se ha cesado de perseguir a los brujos" (Voltaire 1977, 148). Como la semiótica permite comprehender los sentimientos que producen los símbolos y signos, la ética y la filosofía resultan vitales para generar confianza entre negocios, empleados, inversores y clientes.

La relación entre la ética, la moral y la filosofía se podría definir por otra metáfora de tipo computacional, donde la filosofía representaría el ingenio del programador, la moral el código que procesa y muestra las conclusiones, y la ética en la cúspide. Siendo esta última, la noción o pensamiento que otorga sentido a cualquier aplicación informática. La secularización, que comienza a mostrarse con menos complejos desde la Ilustración, es otra consecuencia del Renacimiento, siendo ambas corrientes sociológicas fruto de la necesidad de fortalecimiento de la ética como atributo inherente que fundamenta la condición humana.

Ya avanzado el siglo XX, el concepto ético se expandió hacia nuevos contextos tales como: la ética discursiva, ética deontológica, ética normativa, ética cognitivista o ética universalista. Estas versiones adicionales son consecuencia todas, de discursos totalitarios y genocidas del siglo XIX y XX.

En sincronía con el nacimiento de los Derechos Humanos como institución internacional, nació el Standford Artificial Intelligence Laboratory (SAIL). Un neo-Liceo aristotélico para la inteligencia artificial, robótica, análisis de algoritmos, psiquiatría computarizada, composición artificial artística. Varios años anteriores al aterrizaje del primer humano en la Luna, John McCarthy, precursor de SAIL, abogó

por desarrollar la inteligencia artificial en la exploración hacia Marte. Una parte de la conciencia humana en los años sesenta estaba deseando huir hacia otro planeta donde olvidar la deshumanización, y otra parte humanizarla con nuevas expresiones desde la propia ética. Quienes se han servido de la inteligencia artificial para mejorar la humanidad merecen un homenaje, el compromiso con los demás es un imperativo categórico kantiano, el deber como máxima expresión de virtud, y plenitud vital.

La ética de la compasión o compasiva es una corriente crítica a la ausencia de virtud en la ética ampliada de los años setenta del siglo XX, por la que nos propone regresar a la virtud de la ética que Platón, Aristóteles y Sócrates nos legaron. Esa virtud está estrechamente ligada a la tolerancia entre diferentes, "en la intersubjetividad asimétrica de la víctima frente a la no víctima" (Etxeberria 1998, 66). Curiosamente, el adjetivo artificioso presenta una semántica enriquecida por su transversalidad, que merece un análisis minucioso para comprehender la división de opiniones y actitudes ante la inteligencia artificial. La acepción más contemporánea, en la que la jurisprudencia tiene una relevancia destacada, define artificioso como un objeto o proceso de nula sustancia, incluso un montaje. Sin embargo, previo a la jurisprudencia latina-romana, *artificiōsus*, fue un término apelado por los primeros cristianos para afirmar cualquier creación de alto valor, o virtuoso, tanto en conmemoraciones como para expresar o sentir divinidad. Al llevar a cabo el mismo análisis del vocablo humano, nos encontramos *khamai*, una dicción para aludir simultáneamente, al camaleón y a la tierra. Camaleón como sustantivo de tipo epiceno, ergo, sin género, y tierra como el fruto de la mano o ingenio del hombre. En resumen, la hermenéutica nos asiste para ensamblar lo artificial con el ingenio y la virtud de cualquier humano en su entorno natural. En esta concurrencia la ética de la compasión sirve de baliza donde aplacar los temores de los más reluctantes ante la inteligencia artificial, e incluso del trans-humanismo, o post-humanismo.

Es recomendable que el debate entre la ética sobre la inteligencia artificial medite en la aportación de Francis Bacon. Este autor fue uno de los precursores de la ciencia industrial, un desarrollo fundamental para que la industria anglosajona haya dirigido la economía global entre el siglo XVI y la actualidad. Uno de sus apotegmas o aforismos más metafísicos se refiere al comercio de la mente con las cosas, *commercium mentis et rei.* En realidad "estaba convencido de que, si los hombres querían hacer progresos en la búsqueda de la verdad, debían consultar más la naturaleza que los libros" (Farrington 1971, 2). Este comercio entre la mente humana y la innovación se hace cumbre

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

con el propósito de la inteligencia artificial; por tanto, estamos ante el sueño de Bacon. Empero, este empirista inglés, fue ante todo humanista, cuyo propósito vital fue aplicar la filosofía contemplativa antigua, al bienestar y desarrollo del ser humano.

> Esta otra presunción de que el aprendizaje debería socavar el respeto a las leyes y el gobierno es mera depravación y calumnia sin toda sombra de duda. Es afirmar que un ciego con una guía puede caminar más seguro que un no ciego camina con luz. El aprendizaje hace el pensamiento más amable, generoso, y comprensible al gobierno (Bacon y Thompson 1928, 53).

Aplicando el legado de Bacon a la computación cognitiva, nos significa taxativamente que esta debe aprender cómo es la naturaleza de la mente humana. Una tarea imposible actualmente, puesto que la neurociencia aún se encuentra en proceso de traducir la ontología de la mente humana. La guía que citó Bacon en esta reflexión es homologable perfectamente al algoritmo de cualquier lenguaje de programación. Por tanto, cualquier tipo de inteligencia artificial autónoma o producida por el hombre que se desvíe o ataque cualquier ética válida para la humanidad, habrá aprendido y aplicado lo que le conduce a su final, será, sin duda, una inversión multimillonaria dilapidada.

El bioquímico y pedagogo Luis Miravitlles fue el primer español que comenzó a publicar la inteligencia artificial hacia los años sesenta del siglo XX en televisión española, "crece ya una nueva familia: la de las máquinas de aprender, las máquinas matéticas, o *self-teaching machines*" (Miratvilles 1969, 66). Miravitlles llegó a escribir que en el futuro se produciría una escalada insospechada en las capacidades humanas observando la evolución natural de otros seres vivos a través de la inteligencia artificial. Su razonamiento se apoyó en que el conocimiento ontológico de otras especies, y la relación de estas con su entorno, podrían ser de gran utilidad para mejorar el bienestar de la Humanidad. Es plausible concebir que otra forma de inteligencia o computación cognitiva pudiera facilitar al ser humano atenuar su finitud, pues la compasión no se reconoce como antónimo del optimismo, si como consuelo para quienes se entregan a la hermenéutica que deshumaniza el ingenio del ser humano, separando el ingenio de lo artificial de su propio creador, el ser humano:

> Si la ética es posible en la vida humana, es porque somos finitos, porque no tenemos acceso a los principios, porque el conocimiento

> humano es limitado, porque dudamos, porque no andamos por una camino claro y distinto, porque no alcanzamos verdades firmes y seguras (Mèlich 2013, 78).

2. **Ética y capitalismo**

Adam Smith, fue el precursor del liberalismo, uno de los humanistas más leídos y citados en la historia moderna. La mayoría de escándalos empresariales desde que Smith publicara sus teorías económicas y morales se han relacionado con sus propuestas de libre mercado y la competencia. Sin embargo, en su obra se reconoce la empatía como virtud humana de la filosofía aristotélica, y el imperativo categórico kantiano sobre el deber de respeto al prójimo, como explicita en la cita a continuación:

> Por muy egoísta que se suponga al hombre, es evidente que hay algunos principios en su naturaleza que le interesan por la fortuna de los demás y le hacen necesaria la felicidad de los demás, aunque no obtenga nada de ello excepto el placer de verla. De este tipo es la piedad o compasión, la emoción que sentimos por la miseria de los demás, cuando la vemos o se nos hace concebir de una manera muy viva. Que a menudo derivamos dolor del dolor de los demás es una cuestión de hecho demasiado obvia como para requerir ejemplos que lo demuestren (Smith 1723-1790, 13).

Dos siglos después de Smith apareció la responsabilidad social corporativa (RSC), un paso importante y laudable, empero ineficaz. Como ocurre desde la Grecia clásica, la filosofía nos permitirá comprender los déficits de la RSC. La primera semilla real del capitalismo se fecunda en el siglo XV, más en concreto, en la vindicación a la dignidad humana de Giovanni Pico della Mirandola. La obra de este humanista italiano supuso una revolución ante el vitalismo y tradicionalismo que venía subyugando al ser humano por siglos. Fue la ciencia el contrapeso para superar una divinidad invisible, donde muchos únicamente encontraban desesperanza, enfermedad y miseria (Mirandola 1978). A Pico della Mirandola, a Smith y Voltaire, les secundaron exploradores osados en busca de satisfacer su finitud. Fueron inconformistas o insurrectos, ávidos de satisfacer sus deseos humanos, buscaban escapar de la miseria y depresión vital por falta de oportunidades de no haber nacido bajo un escudo noble. Los mejores socios de estos soñadores fueron mecenas que arriesgaron su propio bienestar, el de sus familias, e incluso su prestigio, esperando la

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

recompensa en modo de materiales preciosos, especies, alimentos, semillas, y hasta capital humano esclavizado. De aquellas aventuras fueron germinando y colapsando las empresas y multinacionales, que necesitaron de centros de intercambio comerciales y financieros, hoy conocidos como mercados de valores, de futuros, permuta de riesgo de crédito, o de opciones.

Tal ha sido el desarrollo, que actualmente se comercia en los mercados financieros internacionales el equivalente al doble de la producción mundial total. Huelga decirlo, pero hubiera sido imposible alcanzar esa cifra sin los avances en tecnología de datos, comunicación y computación. Es cierto que los mercados financieros son mejorables y que la mayoría de sus miembros no ha leído, ni tiene interés por hacerlo, un solo libro sobre ética o filosofía, pero estos mercados terminan siendo implacables con negocios que no generan valor económico y confianza.

Al inicio de la revolución industrial, la esperanza de vida media apenas alcanzaba tres o cuatro décadas para la mayoría de humanos. Sin el capitalismo, es probable que no hubieran surgido el movimiento social, sus sindicatos, la sindicación y el socialismo. Es cierto que las infra-condiciones de vida aún prevalecen para miles de millones de personas, como es cierto que estos seres humanos más desfavorecidos no conocen el mercado libre y el capitalismo, debido a la hegemonía de regímenes políticos corruptos. No son pocos los gobiernos tiránicos que llegan al poder por el idealismo pueril de doctrinas marxistas, o por otras formas aborrecibles, para terminar abusando de su posición de poder, con las que controlan la economía vía monopolios empresariales de capital público, hurtado al pueblo. Miles de millones de seres humanos son víctimas del proteccionismo autárquico, que se prostituye inmediatamente ante multinacionales occidentales sin escrúpulos, para envilecer el legado moral de Smith.

La prudencia y la convivencia se expresan en la pluralidad, otro concepto básico para el éxito de la democracia y el capitalismo. La inteligencia artificial que no acepte o interprete la importancia de la pluralidad no alcanzará el nivel óptimo cognitivo o inteligente, se difuminará en los defectos y vicios en los que la inteligencia humana se muestra en la corrupción moral de los negocios y la política; con esta contundencia se muestra el rigorismo del capitalismo.

La productividad es el alma del capitalismo, porque frena o elimina la inflación o pérdida de poder adquisitivo de las clases menos favorecidas. La inflación representa la mayor fuente de miseria y analfabetismo, como el mayor desafío en los países desfavorecidos, ahogados por la miseria y la desesperanza de sus gentes. La

productividad será la clave para que economía continué produciendo valor económico ante un nivel de población que comienza a dar muestras de estancamiento, o declive demográfico. La inversión en tecnología es el factor más importante para generar productividad, junto a una educación diversa, plural y eficiente. Eduard Punset, compartió una reflexión que recibió personalmente de Dan Gilbert, en la que este psicoanalista le confesó, "la felicidad residía en reconocer nuestras debilidades" (Punset 2007, 325).

El dilema que hoy se concentra en torno a la inteligencia artificial es similar al que generó, y aún ocupa, la energía nuclear. Se trata de una fuente de energía que suministra calor a hogares, hospitales, colegios y universidades a un coste inferior a otras fuentes de energía contaminantes. El ahorro financiero que la energía nuclear aportó, permitió invertir en sanidad, en desarrollar capital humano por una educación más competitiva, mejorar el capital físico por empresas privadas, y en infraestructuras civiles para el transporte de mercancías, de personas o de energía. Vimos que la mejora en la competitividad conduce a mejor productividad, ergo menor inflación, para finalmente, aumentar la renta disponible de hogares y empresas. Con mayor renta, el ahorro y la inversión desde los mercados de capitales o financieros han permitido invertir y desarrollar nuevas fuentes de energía renovables, en sintonía con la salud del planeta y la salud de la humanidad. En la actualidad, la curva de costes de energía muestra dos ganadores ostensibles, la energía fotovoltaica y eólica, y otros que han perdido eficiencia e importancia, como el carbón, petróleo y gas, para desesperanza de dictaduras que se enriquecen vendiendo combustibles contaminantes. Esta curva de coste que se observan en la producción de energías sostenibles y no contaminantes se está acercando a niveles ínfimos, incluso gobiernos hasta no hace mucho anticapitalistas, como el de China, abrazan con jolgorio la inversión en energía solar o eólica. En resumen, la energía nuclear ha favorecido que la humanidad disfrute de fuentes de energía menos contaminantes, y más económicas.

William E. Connolly, filósofo especializado en pluralismo, consigue producir un análisis valioso acerca de la distancia entre la terapia freudiana y la ética. Connolly argumenta que la terapia permite reformar endógenamente aquellos procesos que precisen correcciones como si de algoritmos no automatizados se tratará, frente a los atributos autónomos como la voluntad y la capacidad intelectual inherentes en los humanos (Connolly 2002). Para este autor la ética era una expresión de arte supremo, un reduccionismo para sentir que la ética aplicada es la ruta ideal al éxito. Los algoritmos que desarrollan la

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

inteligencia artificial generativa y sus evoluciones futuras deben integrar el efecto que la ética provoca en el cerebro humano, la dopamina de la felicidad. El éxito de cualquier negocio subyace en la felicidad o comodidad que sus clientes compren. El desafío inédito para las matemáticas en la computación cognitiva, pasa por replicar todos los secretos que el abstraccionismo del alma humana conlleva, humanizarse de forma exógena o endógena (generativa), es la matemática que citó Miravitlles.

Lawrence Kholberg (1958), filósofo y psicólogo, expuso una fórmula excelsa de pedagogía moral. Según su teoría, la mayoría de sujetos morales se clasifican en un primer nivel por su conformismo o vitalismo, tal vez por temor a las represalias, en un sistema donde el dinero es el medio para cubrir necesidades y obtener deseos. En el segundo nivel, se sitúan aquellos humanos que viven desde un enfoque más utilitarista, de compromiso, que demanda dilemas entre las normas recibidas y la moral, presentando mayor autonomía en el procesamiento cognitivo, y protagonismo en sus vidas al imperativo categórica kantiano, por el que debemos regir nuestra moral bajo la regla de oro, o el deber como medio para la conquista del máximo bienestar personal y colectivo. Kholberg expuso también que los individuos con mayor autonomía y autoestima idealista pueden convertirse en los mayores inmorales, por una represión ante la autoridad bajo un patrón que deriva de una relación paternal compleja o no resuelta. El nivel superior de moralidad lo presentan aquellos individuos que planifican, ejecutan y se expresan en términos de impersonalidad, que no se amedrentan ante el castigo, capaces de cultivar el deber y el derecho moral sincrónicamente.

Kohlberg añade una comparación cultural a su análisis, observando niños y adolescentes entre diez y dieciséis años que vivían en EE.UU., Taiwán, México y Turquía. En sus conclusiones se podría inferir que las sociedades urbanas facilitan el tránsito hacia estados morales más autónomos, idealistas y universales, tal vez porque las sociedades urbanas son más permeables a la intersubjetividad, a las experiencias de los demás. Como en su momento el autor no consideró oportuno añadir esta lógica, o no se percató, conviene recordar que EE.UU. vivió un periodo de elevado inconformismo social desde los años cincuenta hasta los años ochenta, ampliado en ámbitos urbanos, con especial impacto en la década en la que Kholberg publicó su análisis. Y, además, que en el caso de Méjico o Turquía, la menor proporción de población urbana pudo reducir la interacción y diversidad social a la secularización más reducida en zonas no urbanas. La investigación de Kholberg es de las más apropiadas a la hora de encargar o contratar

soluciones de computación cognitiva a desarrolladores especialistas, si se pretende evitar fracasos seguros.

3. **Derecho a la privacidad**

Al final de esta sección se incluye un breve análisis práctico sobre el impacto de la falta de respeto al derecho fundamental de la privacidad, en concreto el caso de Meta, antes Facebook. La privacidad y la elección autónoma sobre los límites casuísticos a este derecho de individualidad supone un elemento indispensable de la calidad de vida humana. Como no existe consenso universal sobre una definición de tipo apodíctica para definir concisa y ampliamente el término de privacidad, este ensayo ofrece una revisión desde la perspectiva de dos filósofos especializados en la filosofía de la mente y el pragmatismo. Ambos presentan diferentes inclinaciones sobre propuestas dispares de organización política-dogmática, con argumentos relevantes a la hora de enriquecer la deliberación dialógica que nutran la crítica, pluralidad y tolerancia, valores que se expresan en los Derechos Humanos fundamentales.

Wilfrid Shellars, filosofo norteamericano, realizó parte de sus estudios universitarios en Paris, se mostró renuente sobre el concepto de privacidad absoluta. Shellars afirmó que la única forma de comprender nuestras emociones deriva de la intersubjetividad, sinónimo de convivencia en ética. Analizando a Shellars (1956), se abre una reflexión en torno a las desventajas potenciales en la convivencia de la ética por el anhelo de una privacidad absoluta. Como la privacidad interior toma forma gracias a la intersubjetividad, se podría inferir que la privacidad absoluta o privacidad relevante podrían parecer incompatibles con la empatía, una virtud compartida tanto por las relaciones humanas como la propia ética. La de Shellars es una propuesta sincrónica a dilemas que se observan en la casuística y el principalismo. En este último acercamiento se abrazan con demasiada facilidad reguladores y grupos de influencia a la hora de analizar y resolver otros asuntos y desafíos de carácter socio-económico.

Baruch Spinoza, filósofo, llegó a ser marginado de su propia comunidad hebrea por relacionar a Dios con la Naturaleza; una herejía en su época, pero una relación cuasi integrada hoy en el catolicismo. Esta metáfora de Spinoza sirve para comprender por qué afirmó, en el siglo XVII, que la mente y el cuerpo humano se relacionan desde la unicidad, y que a su vez el cuerpo se nutre de su entorno, bajo un vínculo bidireccional o binario que termina determinando la propia

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

forma de elegir, decidir y razonar de cada individuo. Spinoza concluyó que su propio conocimiento no era suficiente para determinar su concepto de ética; él necesitó comprender la relación de sus coetáneos hacia la ética y el conocimiento. Un pensamiento que puede ser recurrente para definir la intersubjetividad, tal como hoy la conocemos. Spinoza (2019) habló por primera ocasión sobre el respeto a la privacidad en su Tratado Político, y supeditó la piedad de la privacidad a la utilidad pública, porque de colapsar esta última, la ira y la impiedad censurarían la mayoría de espacios de privacidad o derecho a sucedáneos de esta, como el derecho a la propiedad privada. El ingenio excelso de Spinoza invita a preguntarnos hasta dónde la privacidad se convierte en un activo o una carga para el bienestar individual y colectivo de nuestra sociedad, y qué tratamiento se debe requerir a la computación cognitiva en cuanto al sentido de privacidad.

> De lo anterior se sigue que el derecho e institución de la naturaleza, bajo el cual todos nacen y viven la mayor parte de su vida, no prohíbe nada más que lo que nadie desea y nadie puede; pero no se opone a las riñas, ni a los odios, ni a la ira, ni al engaño, ni a absolutamente nada que aconseje el apetito. Nada extraño, ya que la naturaleza no está confinada a las leyes de la razón humana, que tan sólo miran a la verdadera utilidad del hombre y a su conservación, sino que implica infinitas otras, que abarcan el orden eterno de toda la naturaleza, de la que el hombre es una partícula, y por cuya necesidad todos los individuos son determinados a existir y a obrar de cierta manera. Así, pues, si algo nos parece ridículo, absurdo o malo en la naturaleza, se debe a que sólo conocemos parcialmente las cosas e ignoramos, en su mayor parte, el orden y la coherencia de toda la naturaleza, y a que queremos que todas las cosas sean dirigidas según los hábitos de nuestra razón. Pero la verdad es que aquello que la razón define como malo, no es malo en relación al orden y a las leyes de toda la naturaleza, sino únicamente en relación a las leyes de nuestra naturaleza (Spinoza 2019, 267).

Siendo Asia una de las zonas donde la inteligencia artificial ha obtenido mayor aceptación, cabe reconocer que la moral, ética y estilos de vida que predomina en esa zona del planeta son menos sensibles al derecho a la privacidad en Occidente. El confucionismo, una de las corrientes filosóficas más antiguas e influyentes en Asia presenta cierta sincronía con la dignidad del ser humano de Pico della Mirandola, porque asume a las relaciones humanas como el cimiento de su propuesta de pensamiento. La expresión que resume esta filosofía oriental es el *jen-i-li-chih*, (humanismo *-jen-*, fidelidad *-i-*,

propiedad -*li*-, educación en libertad -*chih*). En torno al confucionismo gira la creencia de que, tarde o temprano, cualquier sujeto necesitará de la generosidad y empatía de otros humanos de su entorno, por tanto, no es inhumano anteponer la intersubjetividad de las relaciones personales ante la autonomía individual (Yum 1988). En Asia, es más común encontrar organizaciones económicas más horizontales que dependen más de la comunicación y la confianza que las verticales; en estas últimas, la privacidad se reconoce y afirma.

La computación cognitiva es una consecuencia de la inteligencia artificial, y la primera será reemplazada por otro desarrollo superior en ingenio. Según Dirk Helbing (2015), subestimar la importancia de estas evoluciones supondría una suerte de lotería de consecuencias impredecibles. El negacionismo digital o vitalismo, basado en el temor ante una inteligencia más eficiente, conlleva otro riesgo igual o tal vez superior a una inteligencia no humana, una pendiente resbaladiza por el desconocimiento que erigiría otra división en las capacidades y oportunidades al alcance de cualquier humano.

Ya que la computación cognitiva se alimenta de aprendizaje de tipo auto-generativo, automático y cognoscitivo, cualquier proceso de supervisión debería estar abierto a nuevos algoritmos específicos de control endógeno o exógeno. Es prudente instruir a los algoritmos la capacidad de interactuar entre ellos, para supervisarse mutuamente con atribuciones de justicia y coercitivas. La práctica de moralidad o justicia entre algoritmos formará parte de su intersubjetividad digital; la inteligencia humana debe ser policía y fiscal dentro de este mecanismo de control, de esta forma se detectará cuando un algoritmo se corrompa por interés propios, con la aquiescencia de otros algoritmos que se verán seducidos por manipular su diligencia debida, en modo similar al tráfico de influencias o intercambio de favores que es común en oligopolios desde hace siglos. Sería cándido asumir que la computación cognitiva no será tentada de auto corromperse. La corrupción en la computación cognitiva es más que plausible, en adelante analizaremos las neuronas espejo para tratar de asumir este riesgo.

Hace tres décadas, el neurólogo Giacomo Rizzolatti presentó su descubrimiento sobre las neuronas especulares, o neuronas espejo. Son neuronas esenciales para la empatía que se activan para reconocer las intenciones del resto de sujetos con quienes nos relacionamos. Se reconocen como las neuronas responsables del aprendizaje del lenguaje, sin ellas la comunicación humana no sería tal como la conocemos, a buen seguro la evolución humana presentaría otra historia muy diferente (Martín Loeches et al. 2008). La privacidad es

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

como un velo de ignorancia de nuestras intenciones en los demás, que protege la intimidad, pero no sería prudente olvidar que la comunicación, el lenguaje y la empatía no se nutren de la privacidad, al menos sin conocer nuestros pensamientos. Un exceso de privacidad, podría derivar consecuencias tan graves en la estructura social de cualquier comunidad y en su convivencia, como la falta de intimidad absoluta. De ahí que no exista una definición unívoca y universal del concepto como tal de privacidad.

La convivencia como base de la privacidad y fundamento ético invita a reflexionar en cómo facilitar la ontogenia de la computación cognitiva. El psicólogo y filósofo Henry Ey (1967) se ocupó de analizar la herencia del pensamiento, y los neurotransmisores, enfocando las teorías de la personalidad a un base elemental, donde su fisionomía (huellas digitales) se fundamenta de forma exógena. Esta reducción ilustra la fase en la que la computación cognitiva se encuentra actualmente, siendo este factor exógeno el reflejo del medio, es decir, la influencia de la moral y las características éticas o sesgos emocionales de los desarrolladores humanos y de quiénes a su vez, supervisan a estos con el objetivo focalizado en los resultados financieros. Las propuestas de computación cognitiva o humanizadas de mayor éxito en el futuro serán aquellas que faciliten la comprensión de los beneficios que supone la privacidad y, sobre todo, las soluciones capaces de adaptarse de forma más eficiente a las diversas nociones de privacidad en su uso y limitaciones, a través de todas las culturas diversas en el mundo.

El Super-Yo artificial es la fase posterior a la computación cognitiva, que podríamos comenzar a definir como computación sub-consciente e inconsciente. De forma autónoma, podría llegar a ser capaz de comprender la dimensión de la privacidad como un rasgo que forma parte de los derechos fundamentales de los seres humanos. En la investigación de Freud sobre el narcisismo, encontramos esta expresión: "*Wo es war soll ich werden*", traducido al castellano, "allí donde era Ello, debo hacerme Yo" (Freud et al. 1973). El deseo y anhelo de autonomía individual se enfrenta a esta moral común aceptada, y provoca sentimientos enfrentados que suponen un riesgo para la convivencia con uno mismo y la comunidad. Por supuesto que no es asumible trasladar la biología humana a la ontología incipiente de la inteligencia artificial, pero las circunstancias inciden en el desarrollo de la personalidad humana, y la inteligencia artificial aún no es plenamente autónoma de la inteligencia humana. Para cualquier empresa que pretenda capturar la demanda de soluciones de inteligencia artificial, le será imperativo reducir las consecuencias del

potencial complejo de edipo de la computación cognitiva. El axioma de los riesgos es la pérdida de confianza ante usuarios, clientes e inversores, y la amenaza de este riesgo es la capacidad o influencia colectiva que expone Young:

> Sindicatos, grupos religiosos, y organizaciones de accionistas son algunas de estas entidades que a veces ejercen un poder significativo, no porque puedan coaccionar a otros para imponer sus decisiones, sino porque cuentan con muchos miembros que actúan juntos […] existe una demanda moral altisonante […] Una actitud responsable de aquellos con un interés fundamental en acabar con las injusticias no es culpar a los poderosos […] sino responsabilizarlos públicamente (Young 2011, 154, 156)

Las redes sociales han multiplicado la audiencia en todo el mundo, acelerando el acceso a la información y la influencia de opinión, el riesgo de prestigio o credibilidad se ha convertido en una de las mayores amenazas para el capitalismo e inversores, mayor incluso a los riesgos del clima, económicos o financieros. Para comprender mejor las consecuencias de menospreciar esta contingencia, conviene analizar el caso de Facebook, hoy Meta. El 10 de abril de 2018, Mark Elliot Zuckerberg, uno de los fundadores de Facebook, hoy Meta, compareció ante el Congreso de EE.UU., para rogar por el perdón de los ciudadanos del mundo y de los representantes del poder legislativo en ese país. Las primeras sospechas y denuncias contra Facebook por violar la privacidad de sus usuarios comenzaron a ser públicas hacia el año 2008, sin embargo, a finales del 2016, se desató el escándalo más grave en materia de privacidad en la última década. Se constató que Facebook vendió datos de millones de usuarios sin su visto bueno y conocimiento a una consultora británica que analiza las intenciones de voto de ciudadanos (NBCNews 2018).

Es elocuente observar el valor de mercado de todo el negocio de Facebook o Meta durante la última década, ese valor es el interés de los inversores capitalistas, que depende de la confianza que esta empresa produce. En las primeras columnas de la tabla siguiente se observa el precio de cotización al final de cada año natural desde el final de 2013, para las acciones de Meta-Facebook, Netflix y Amazon, corporaciones cuyos modelos de negocio dependen en buena medida de la minería de datos, de su interpretación y distribución. En las tres últimas columnas se observa la revaloración acumulada de las acciones de estas tres empresas al final de cada año natural. Sincrónicamente, el interés de los inversores en acciones de Facebook se debilitó a partir del testimonio de

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

Zuckerberg en el Congreso de EE.UU., mientras creció el interés de inversores en las acciones de Netflix y Amazon. La variable más importante para inversores reside en la confianza que pueden depositar en la moral de los directivos de las empresas, ya que a largo plazo determina el éxito o colapso de cualquier negocio. Ese carácter se observa en la empatía de la cultura de la corporación ante clientes, empleados, inversores, proveedores, reguladores. Cuando la empatía de una organización decae, se eleva la incertidumbre y la sensación de riesgo, empujando al alza la prima de rentabilidad que se utiliza para valorar en el presente los flujos de beneficios a producir en el futuro.

Tabla 1

Evolución bursátil. 2013-2023: META, AMAZON, NETFLIX

	META	AMAZON	NETFLIX	META	AMAZON	NETFLIX
	Precio acciones			Revalorización del valor de mercado acumulado		
27/12/2013	$55,44	$19,90	$52,50			
26/12/2014	$80,78	$15,45	$48,58	145,71%	77,64%	92,53%
31/12/2015	$104,66	$33,79	$114,38	188,78%	169,80%	217,87%
30/12/2016	$115,05	$37,49	$123,80	207,52%	188,39%	235,81%
29/12/2017	$176,46	$58,47	$191,96	318,29%	293,82%	365,64%
28/12/2018	$133,20	$73,90	$256,08	240,26%	371,36%	487,77%
27/12/2019	$208,10	$93,49	$329,09	375,36%	469,80%	626,84%
31/12/2020	$273,16	$162,85	$540,73	492,71%	818,34%	1029,96%
31/12/2021	$336,35	$166,72	$602,44	606,69%	837,79%	1147,50%
30/12/2022	$120,34	$84,00	$294,88	217,06%	422,11%	561,68%
29/12/2023	$353,96	$151,94	$486,88	638,46%	763,52%	927,39%

Fuente: Elaboración propia a partir de Google Finance

Con estos datos, es plausible estimar el coste de oportunidad para los accionistas o inversores de Meta, en su día Facebook, por la escasa moral y ética de esta empresa. Sin escándalos por medio, se podría inferir que la apreciación en la acción de Facebook al cierre de 2023, presentaría evolución similar a la de los títulos de Netflix o Amazon

durante 2013-2023. Tomando una media geométrica simple, la acción de Meta podría haber ascendido a $460 a cierre de 2023. Amazon también ha recibido sanciones por violar normas de privacidad, pero en bastante menor escala a las que recibió Meta-Facebook. Según los últimos estados contables y financieros registrados por Meta ante el regulador de los mercados financieros en EE.UU. el beneficio tras impuestos para el ejercicio último, alcanzó $14,87 por cada acción en circulación (SEC 2024).

Por tanto, es plausible concluir que el coste de oportunidad para los accionistas de Facebook equivale a $460 – $353,9, donde $353,9 es la cotización real a cierre de 2023, y $460 es la cotización potencial de acuerdo a la evolución del interés de los mercados en Netflix y Amazon. Este coste de oportunidad representa más de siete veces el beneficio neto del último ejercicio, una cifra desorbitante. Así de implacables son los mercados financieros ante la falta de ética y moral en las empresas, por esto recalcar que Adam Smith, el precursor del capitalismo, fue filósofo, sobre todo, y luego economista.

4. **Derecho al trabajo**

El artículo 23 de la Declaración Universal de los Derechos Humanos incumbe a dos aspectos humanos cruciales, el derecho de acceso y equidad en las oportunidades laborales. La historia económica exhibe que las disrupciones tecnológicas han deslocalizado y transformado las funciones del capital humano con una facilidad asombrosa, por lo que es razonable asumir dos niveles de consecuencias. Por un lado, cientos de millones de empleos se extinguirán; para tratar de calibrar este impacto, se pueden distinguir tareas por los atributos siguientes: rutinario o no rutinario, cognitivo o no cognitivo, de elevada o reducida especialización (Ted Tschang y Almirall 2021). Por otro lado, nuevos puestos de trabajo surgirán, como la productividad aumentará, será posible reducir la jornada laboral, mantener o aumentar la renta disponible de hogares, y la recaudación fiscal, para ofrecer las políticas sociales más relevantes. Esta es la fórmula que ha llevado la humanidad de la Edad Media a la vida actual, con todo lo que queda por mejorar, por supuesto.

En sintonía con la ética de la compasión, y la finitud del ser humano, la doctora en sociología Helga Nowotny, nos exhorta a reflexionar ante los prejuicios que derivan de la finitud humana:

> El miedo a nuevo inhibe la curiosidad, incluso cuando no se deja intimidar del todo por el miedo que ella misma provoca. Pero la

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

> curiosidad comienza a vacilar y tambalea […]. El polo opuesto, el subjetivo, abarca todo sentido, belleza, significado, y valor que le atribuimos a la realidad objetiva. Pero esta separación, que todavía sigue estando muy arraigada en el pensamiento occidental, en la filosofía, y en lo que aún no dijeron las otras disciplinas, ignora que no hay ninguna diferencia de base entre nuestro acceso al conocimiento y nuestra manera de conocer. (Nowotny 2024, 45, 46)

Para Nowotny, la filosofía es indispensable para comprender la razón humana, y la ciencia incluida la inteligencia artificial son herramientas desarrolladas al servicio de los humanos. Esto que parece tan obvio observando la historia, no se debe olvidar por quien promueva la tecnología, para evitar que la opinión medieval más fundamentalista acabe con la Ilustración y el Renacimiento. Esta última corriente más contraria al desarrollo técnico se presenta en una suerte de falacia naturalista para reafirmar que el ser humano es lo que debe ser, libre de amenazas potenciales como la robótica, o la inteligencia artificial generativa. "El vitalista moral asegura que lo debido es cuanto colabora con la vida para magnificarla y asegurarla, siendo indebido y malo cuanto la compromete y la desmiente" (Savater 1992, 298).

La corriente filosófica taoísta resulta eficaz para reivindicar un tipo de liderazgo que destaque por: su disposición por el conocimiento, la formación constante, determinación, y por otros atributos de los directivos que determinan el compromiso de los empleados en la organización: la benevolencia, su empatía con el equipo, y su elevado sentido de justicia (Bai y Roberts 2011). Estos últimos atributos presentan una notable correlación con el nivel moral post-convencional que ya se ha citado previamente, en la teoría del desarrollo moral de Kholberg.

Aplicando otro caso práctico para este debate, en concreto el de José López Rojas, doctor en finanzas, y especialista en Python, una de las aplicaciones de mayor ocupación en la inteligencia artificial. "Las finanzas del siglo XXI han evolucionado significativamente gracias a los avances tecnológicos, particularmente en el ámbito de la inteligencia artificial (IA). Estos avances han transformado cómo las instituciones financieras operan, toman decisiones y sirven a sus clientes" (López Rojas 2024a, 2). Python facilita procesos de análisis bajo un entorno macro y micro económico dinámico, influido por millones de variables cuantitativas y cualitativas no estacionales, por las que las conclusiones del algoritmo deriven de acontecimientos des-correlacionados con el pasado. Incluso, sería factible contemplar el escenario por el que las variables seleccionadas estuvieran correlacionadas, un efecto que se

conoce como multicolinealidad estadística. Si el emisor de estos activos financieros fuera cualquier multinacional con balances multimillonarios que hubiera colapsado sorprendentemente por un escándalo de corrupción o similar, el algoritmo debería estar programado para predecir las posibilidades de bancarrota de sus competidores, como también de otros riesgos económicos o financieros de tipo sistémico.

Es tal el nivel de complejidad por la diversidad de información, conocimiento, y disciplinas, que se haría necesaria la intervención de la inteligencia humana con nuevas funciones adicionales a las que hasta ahora han desarrollado los analistas financieros, como la valoración de los riesgos operativos y financieros por ausencia de suficiente diligencia ética ante determinadas decisiones de los directivos de ese negocio o la competencia. Nuevos perfiles laborales con competencias inéditas y capacidad de lógica multidisciplinares serán necesarios, así como diseñar programas de estudio universitario o de formación profesional capaces de anticiparse a los nuevos retos, para proveer un capital humano más renacentista, diverso.

> La innovación y la ética deben ir de la mano en este nuevo escenario, pero, sin una dosis de creatividad y aporte único, el desarrollo humano y el crecimiento profesional corren el riesgo de estancarse. Este cambio de paradigma no admite términos medios: o los trabajadores se mantienen evolucionando constantemente o se quedarán atrás (López Rojas 2024b, 8)

Esta cavilación contemporánea no dista del pensamiento de Pico della Mirandola, puesto que la inteligencia artificial como desarrollo científico no se debe observar únicamente como una amenaza, sino también como una palanca para expandir la dignidad humana por la responsabilidad inalienable como contraprestación a la virtud heredada de nuestros antepasados.

Seyla Benhabib, filósofa especializada en política, afirma que la vida social y económica que fundamenta la intersubjetividad depende de dos pilares, la cooperación y de aceptar las consecuencias impredecibles de nuestras acciones o decisiones (Benhabib 2004). La mayoría de los empleados vulnerables a la computación cognitiva se verán empujados a un inevitable proceso de evolución y adaptación, como el camaleón que sirvió a los griegos antiguos para construir la palabra humano de su época que nos legaron. Este desplazamiento laboral es comparable en la distancia con las necesidades que los migrantes presentan cuando se ven abocados a un traslado a cientos o miles de kilómetros de distancia. Benhabib propone aceptar una forma

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

de redistribución económica al estilo que proponen otros filósofos como Thomas Pogge o Charles R. Beitz, pero siempre que esta generosidad económica se subordine a los preceptos siguientes

— Objeciones epistémicas-hermenéuticas: ¿Todos los puestos de trabajo que se eliminen en el futuro serán una consecuencia de la computación cognitiva? Según el índice de desarrollo humano, que publica el programa de desarrollo de la ONU, durante las tres últimas décadas la calidad de vida ha mejorado en general, un 25% en todo el Mundo (Naciones Unidas 2024). Desde la polea de Arquímedes, la tecnología es una respuesta a la búsqueda de eficiencia económica, sanitaria, social, como consecuencia del ingenio innato inalienable a la condición humana y a la insatisfacción de su finitud, pues no ha cesado de la búsqueda por la innovación, con el doble fin de reducir el esfuerzo humano o aumentar los resultados que la ciencia económica reconoce como productividad. Sin el desarrollo de la ciencia y la técnica, los flujos migratorios actuales no tendrían sentido, ya que la forma de vida sería la misma que hace miles de años para todos por igual, el ser humano no habría conocido el privilegio de la universidad, de la atención hospitalaria, del placer de la cultura o el turismo. La industria del turismo supone el 10% de la Economía mundial actualmente, es la industria con mayor crecimiento en la última década. Con el aumento de la productividad de la tecnología, se reducirá aún más la jornada laboral, aumentando el tiempo de ocio simétricamente; por tanto, esta industria demandará más capital humano.

Benhabib propone más derechos, pero también mayor transparencia, y sentido del deber, para que cada ciudadano pueda y perciba la necesidad de reflexionar, reconocer y afirmar un consentimiento informado de los cambios en el mercado laboral que la computación cognitiva provocará, así como de las alternativas que dispone para beneficiarse de esta inercia, o las ventajas y riesgos que asumirán de optar por mantenerse al margen. Para quienes prefieran rechazar esta corriente por su balance de obligaciones y derechos comunitarios e individuales, es prudente mantener el sentido de justicia distributiva, recordemos que la ética es prudencia más convivencia o amistad.

— Objeciones democráticas: Bajo una perspectiva más racional o económica, una de las condiciones para que cualquier tecnología se desarrolle y perviva es que reduzca, o no amplie,

> la distancia de bienestar y calidad de vida entre clases sociales, como se muestra en el índice de la ONU que se referencia previamente. Es imperativo que no se censure directamente el acceso a los beneficios de la computación cognitiva para ningún ser humano, aunque no es verosímil plantear este derecho simétricamente, dado la diversidad de culturas y de principios morales en una sociedad global plural. Benhabib denuncia lo injusto de los privilegios que la ONU aún permite a un puñado de países, en detrimento de los intereses del resto; traducido a un contexto corporativo, estaríamos observando un cartel u oligopolio.

La computación cognitiva debe ser útil al ser humano para impulsar la competencia en el sector tecnológico, reduciendo la influencia de pocas compañías en la dirección de esta industria, que supone una autarquía capitalista oligopolista. La democracia es el fundamento de Smith y de su mercado libre; en la búsqueda de la autodeterminación individual es necesario aumentar el rango de oportunidades, se trata de ampliar al máximo las oportunidades de trabajo para la mayoría, y evitar concentrar la demanda y oferta de talento en un número reducido de organizaciones, lo que provocaría la frustración y desactivación de buena parte del talento humano. Por esto mismo, la pedagogía en la inteligencia humana, como medio para el fin de convivir con la computación cognitiva, debe abrirse a nuevas perspectivas además de la lógica o la razón, otorgando mayor relevancia, por ejemplo, al humanismo y los valores democráticos y de justicia. Esta recomendación debe viajar más allá de los centros educativos para jóvenes o niños, hasta la formación continuada en lugares de trabajo de cualquier índole.

Meta-Facebook fue desde 2004 un nirvana para los programadores e ingenieros de computación que terminaban su carrera universitaria. Ver años después al fundador reconocer desde el Congreso los fraudes y abusos cometidos desde la organización de la que depende tu bienestar financiero y el de tu familia, debió provocar cierta angustia entre directivos y empleados de esta empresa. Desde la perspectiva moral, un buen número de empleados en Facebook debió sentir cierta incomodidad por la displicencia de la compañía ante su deber con las necesidades y derechos del resto de personas que dependían directa o indirectamente de la estrategia de negocio corporativa. Siento estos empleados, además, usuarios de las soluciones de Facebook, por lo que sus propios datos fueron utilizados de forma fraudulenta. No sorprende el éxodo de un buen número de profesionales y directivos

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

de Meta-Facebook, con el consiguiente riesgo que estas dimisiones representaron para la estabilidad y solvencia de la compañía (Franck 2019). Quienes optaron por abandonar ese proyecto se unieron a la competencia o crearon su propio negocio, integrando toda la experiencia y conocimiento acumulado financiado con el capital y la confianza de los accionistas de Meta-Facebook.

Conclusiones

En sintonía con el argumento que aplica Francisco José Blanco Brotons en su crítica a la teoría de la justicia de Rainer Forst: "Una propuesta normativa basada en la capacidad de proclamar las razones «no refutables» cae en un idealismo racionalista incongruente con la teoría discursiva" (Blanco Brotons 2023, 266). Desde la diversidad cultural y moral de la que disfruta la humanidad, como la prudencia y convivencia que aporta la ética, no cabe abordar los retos de la computación cognitiva para los derechos humanos fundamentales desde una perspectiva legal apodíctica.

Investigadores japoneses acaban de publicar una nueva tecnología basada en inteligencia artificial que es capaz de representar en imágenes la actividad neuronal profunda de la mente humana. Este descubrimiento representa un hito en la investigación neuronal que aumenta la competencia en investigaciones similares que nos trasladarán a contextos de intersubjetividad desconocidos por el ser humano, para desafiar los derechos humanos fundamentales conquistados hasta ahora (Koide-Majima et al. 2024). No debería sorprendernos que en un futuro no lejano se pueda reconstruir o mapear nuestro pensamiento para evitar situaciones de extrema violencia en zonas de transporte público, que las aplicaciones sociales de contacto actualicen sus versiones para conocer las emociones que producimos en otras personas, o que los equipos de recursos humanos puedan valorar la actitud de sus nuevas incorporaciones.

Ante el rechazo o contradicción inherente en la ontología de cada humano que este tipo de avances suponen para respetar derechos fundamentales como la privacidad, la ética se erige *per-se* en el principal factor diferenciador. Los inversores o empresarios más avezados comprenderán que este rechazo deriva de una necesidad básica de consumidores por proteger su autonomía y dignidad personal. Los planes de negocio que pretendan posicionarse con ventaja respecto a la competencia para atender de manera más ética y eficiente la relación de los consumidores con su privacidad harían bien en combinar

las teorías de estrategia de negocio como las de Michael E. Porter con los derechos humanos fundamentales. Este economista desarrolló una teoría que permite reconocer las ventajas competitivas de cualquier negocio, en relación a su posición ante clientes, proveedores, la posibilidad de aumentar la competencia en el sector, de nuevos productos o soluciones, y del nivel de competencia (Porter 1979).

Cuando Porter publicó esta teoría, la inteligencia artificial se encontraba en una fase muy incipiente, cuarenta años después está transformando los modelos de negocio de empresas con décadas de trayectoria, y dando paso a nuevos modelos de negocio que podrían canibalizar la cuota de mercado a corporaciones hoy muy rentables y atractivas para inversores. En los últimos años se ha desarrollado la cuarta revolución industrial, que reconoce y afirma contextos de organización corporativa vanguardistas, cuyo objetivo básico es aumentar la productividad, y las economías de escala. La cuarta revolución industrial es compatible únicamente, con modelos de negocio que demuestren una capacidad de adaptación excelsa, mediante procesos de aprendizaje automáticos, aplicando y desarrollando algoritmos de última generación (Lanteri 2021).

Desde su metáfora de los mercados financieros, Benhabib sugiere que las prácticas inmorales deben ser perseguidas y condenadas, porque ponen en riesgo el funcionamiento eficiente del sistema de cooperación, es decir, nuestra confianza en los demás. Subraya también que nuestras decisiones y acciones influyen en el bienestar de los demás, afirmando el concepto de responsabilidad moral como el eje de su análisis, puro kantianismo. En correlación con esta metáfora, es posible inferir que aquellas aplicaciones de inteligencia artificial capaces de ubicar los derechos humanos fundamentales como medio y fin a su vez en la estrategia de negocio, estarán en situación de ventaja competitiva ante sus competidores. La confianza está muy correlacionada con la convivencia; este último atributo es un rasgo esencial de la ética, y los usuarios o consumidores optarán por aquellos desarrollos que ofrezcan mayor empatía y respeto por el interés personal de estos. Sin olvidar que la multiculturalidad invita a considerar la diversidad en la relación de los humanos con determinados derechos fundamentales, ya se ha citado que bajo el confucionismo, la privacidad podría generar sentimientos diferentes entre ciudadanos de cultura japonesa, coreana, china, con respecto a los influidos por culturas occidentales.

Smith aportó una teoría moral excelsa, pero muchas corporaciones no han interpretado y aplicado sus cánones morales, que se podría resumir en que la confianza debe ser la principal ventaja competitiva en

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

el plan de negocio de cualquier empresa (Shleifer 2004). Tampoco lo hizo el propio Porter al inicio de su carrera académica. Como el vitalismo es parte de la condición humana, es lógico que alguno de los lectores de este documento piense de inmediato que la ética reduce las posibilidades de negocio olvidando la recomendación de Punset y Gilbert sobre el reconocimiento a las debilidades del ser humano, y su auto impotencia para el éxito y la felicidad que se expone a lo largo de este documento. El valor social y económico de cualquier tecnología deriva de la confianza, y esta a su vez del respeto por la moral. La humanidad siempre marginó antes o después aquello que produce desconfianza, que no respeta los códigos de moral necesarios. Son conocidas las consecuencias que produce el *doping* en la reputación e ingresos de los deportistas profesionales que no respetan el juego limpio. De la misma forma, el algoritmo que no respete la moral acabará expulsado. Aristóteles definía como *idiotés* a los griegos que se auto marginaban de la polis.

La cuarta revolución industrial dispersa la información y los datos como el oxígeno, los consumidores y empleados adquieren mayor sentido de su poder para marginar negocios que no ofrecen servicios de valor económico-social, desde el respeto a los derechos humanos fundamentales. El futuro de los negocios y la evolución de la conciencia humana exige a las empresas adaptar su estrategia y organización constantemente, hacia un talento humano que ya no se conforma con una salario digno y atractivo, exige además sentirse alineado con la cultura del negocio, sentir bienestar en su salud física y emocional.

En los mercados financieros, donde se socava o fortalece el buen hacer de las empresas, y cuyo éxito es a su vez una función más directa de cómo se reconoce la ética y cultura de las empresas, se desarrolla con velocidad un nuevo concepto de inversor, que se conoce como inversor activista, e inconformista. El propósito básico de estos inversores cada vez no tan singulares, consiste en forzar la transformación o adaptación del plan estratégico de cualquier negocio donde decidan invertir, hacia un propósito en sintonía con la propuesta de Smith, ser más empático con el entorno y con los seres humanos en general (Alden 2011). Desde una perspectiva de largo plazo, esta decisión es la más prudente e inteligente para acceder al mejor capital humano posible. Son las personas quienes consiguen el éxito en cualquier aventura o negocio, ellas definen la calidad en los productos o servicios y, por tanto, el interés de consumidores, clientes e inversores, y detrás de estas personas que integran el capital humano, se encuentran sus necesidades, ilusiones, esperanzas, temores, es decir, su propia filosofía.

La responsabilidad social corporativa es débil, apenas escucha la sabiduría de la filosofía. Se fundamenta básicamente en la perspectiva del principalismo y la deontología, en detrimento de la diversidad casuística caracterizada en la personalidad y las necesidades dinámicas e individuales del resto de agentes, como empleados, clientes, inversores, acreedores. Armand D'Angour (2022) nos evoca la batalla de Leucra, (371 a.C.) un acontecimiento único para comprender por qué la filosofía y la ética aportan un valor incomprendido e infravalorado en los negocios. Epaminondas, el comandante de los tebanos, representó la virtud ecléctica de Aristóteles, por su creatividad, inteligencia emocional, reflexión y la anticipación (D'Angour 2022). Su triunfo supone uno de los axiomas de disciplina, trabajo en equipo y compromiso más memorables en la antigüedad. Como el éxito se multiplica en un círculo virtuoso, debido a la audacia virtuosa de Epaminondas, los espartanos se vieron en desventaja al contemplar a su líder Cleómbotro malherido, quien en lugar de ordenar a los soldados abandonarle en su ocaso, les expuso a perecer juntos y perder la batalla. Cleómbotro, anteponiendo su propio interés al ajeno, incumplió el imperativo categórico kantiano, por el que el deber con los demás significa el éxito real, el único. Aquel episodio fue el inicio del declive en la reputación e influencia de Esparta, en la Hélade.

Trasladando las lecciones de aquella batalla a los negocios de la cuarta revolución industrial, la alternativa más inteligente para reclutar y educar capital humano es observar minuciosamente el análisis de Kholberg sobre el desarrollo moral, en concreto la fase post-convencional. Estos empleados piensan y actúan en sentido contrario a Cléombotro, y se simpatizarían con Epaminondas, para construir y gestionar una organización suficientemente atractiva, capaz de atraer más talento, diferenciarse de la competencia, y producir valor económico con menor riesgo. El éxito de cualquier tecnología desde la invención del fuego dependió y dependerá en exclusiva en su capacidad por contribuir a engrandecer el aspecto personal de la existencia humana, lejos del abuso de poder o privilegios injustos (Mitcham 1989). En la era de los datos y la información al instante será muy fácil para cliente y empleados calibrar la honestidad y compromiso social de los negocios.

Referencias bibliográficas

Alden, Eric. 2011. «Shareholder activism by public pension funds and the rights of dissenting employees under the First Amendment», *Harvard Journal of Law & Public Policy* 34: 289-366.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

Aristóteles. n.d. *Ética a Nicómaco.* Barcelona: Orbis.
Bacon, Francis y Matthew Thompson. 1928. *Selections*. Acceso el 1 de Octubre de 2024: http://books.google.com/books?id=l4c3AAAAMAAJ
Bai , Xuezhu y William Roberts. 2011. «Taoism and its model of traits of successful leaders», *Journal of Management Development* 30 (7/8): 724-739.
Benhabib, Seyla. 2004. *The rights of others: Aliens, residents, and citizens*. Cambridge: Cambridge University Press.
Blanco Brotons, Francisco José. 2023. «El idealismo discursivo en la teoría de la justicia: una crítica a Rainer Forst». *Cuadernos salmantinos de filosofía* 50: 247-268.
Connolly, William E. 2002. *Neuropolitics: Thinking, culture, speed*, NED-New edition 23. Acceso el 16 de abril de 2024. http://www.jstor.org/stable/10.5749/j.cttts8p6.
D´Angour, Armand. 2022. *Aristóteles, el arte de innovar.* Badalona: Kõan.
Etxeberria, Xabier. 1998. *Ética básica.* Bilbao: Universidad de Deusto.
Ey, Henri. 1967. *La conciencia.* Madrid: Gredos.
Farrington, Benjamin. 1971. *Francis Bacon, filósofo de la Revolución Industrial.* Madrid: Ayuso.
Franck, Thomas. 2019. «Facebook is downgraded as analyst warns its executive exodus could be contagious». *CNBC*, marzo 18.
Freud, Sigmund, Ramon Rey Ardid y Luis López-Ballesteros. 1973. *El yo y el ello.* Madrid: Alianza.
Helbing, Dirk. 2015. «Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies», *Arxiv*. Doi: 10.48550/arXiv.1504.03751
Kholberg, Lawrence. 1958. *Kohlberg's original study of moral development («The development of modes of thinking and choices in years 10 to 16»).* New York: Garland Publishing.
Koide-Majima, Naoko, Shinji Nishimoto y Kei Majima. 2024. «Mental image reconstruction from human brain activity: Neural decoding of mental imagery via deep neural network-based Bayesian estimation.» *Neural Networks* 170: 349-363.
Lanteri, Alessandro. 2021. «Strategic drivers for the fourth industrial revolution», *Thunderbird International Business Review* 63 (3): 273-283. https://doi.org/10.1002/tie.22196
López Rojas, José. 2024a. «Inteligencia artificial: Reconfigurando el paradigma financiero del siglo XXI», *III Encuentro Nacional e Internacional de Contabilidad y Fiscalidad.* Rionegro, Antioquía. 8.
López Rojas, José. 2024b. «Tras la llegada de la IA, ¿caerá el invierno sobre el mercado de trabajo?». *The Conversation*, 30 de abril.
Martín Loeches, Manuel, Pilar Casado y Alejandra Sel. 2008. «La evolución del cerebro en el género Homo: la neurobiología que nos hace diferentes». *Revista de Neurología* 46 (12): 731-741.
Mèlich, Joan-Carles. 2013. *Ética de la compasión.* Barcelona: Herder.
Mirandola, Giovanni Pico della. 1978. *Discurso sobre la dignidad del hombre.* Buenos Aires: Goncourt.

Miratvilles, Luis. 1969. *Visado para el futuro.* Madrid: Salvat.
Mitcham, Carl. 1989. *¿Qué es la filosofía de la tecnología?* Barcelona: Antrophos.
Naciones Unidas. 2024. *Human Development Index.* Acceso el 12 de mayo de 2024. https://hdr.undorg/.
NBCNews. 2018. *'I'm sorry': Facebook CEO Mark Zuckerberg delivers opening statement at senate hearing*. abril 18. Acceso el 15 de mayo de 2024. https://www.youtube.com/watch?v=UofMQ8EGmSc.
Nowotny, Helga. 2014. *La curiosidad insaciable*. Barcelona: UOC. Acceso el 15 de mayo de 2024. https://lectura.unebook.es/viewer/9788490291122
Porter, Michael E. 1979. «How competitive forces shape strategy», *Harvard Business Review* 57 (2): 137–145.
Punset, Eduard. 2007. *El alma está en el cerebro.* Barcelona: Destino.
Savater, Fernando. 1992. «Vitalismo», en *Concepciones de la ética,* editado por Victoria Camps, Osvaldo Guariglia y Fernando Salmerón, 297-308. Madrid: Trotta.
SEC. 2024. *United States. Securities and exchange commission.* abril 24. Acceso el 12 de mayo de 2024. https://www.sec.gov/ix?doc=/Archives/edgar/data/1326801/000132680124000012/meta-20231231.htm.
Sellars, Wilfrid. 1956. *Empiricism and the philosophy of mind.* Cambridge Massachusetts: Harvard University Press.
Shleifer, Andrei. 2004. «Does competition destroy ethical behavior?.» *American economic review* 94 (2): 414-418.
Smith, Adam. 1723-1790. The theory of moral sentiments / Adam Smith ; Introduction by Amartya Sen ; edited with notes by Ryan Patrick Hanley. New York: Penguin Group.
Spinoza, Benedictus de. 2019. *Tratado teológico-político*. Madrid: Verbum. Acceso el 12 de mayo de 2024. http://www.digitaliapublishing.com/a/63862/.
Ted Tschang, Feichin y Esteve Almirall. 2021. «Artificial intelligence as augmenting automation: Implications for employment». *Academy of Management Perspectives,* 35 (4): 642-659.
Voltaire. 1977. *Tratado de la Tolerancia.* Barcelona: Crítica.
Young, Iris Marius. 2011. *Responsabilidad por la justicia.* Madrid: Morata. Paideia Galiza Fundación.
Yum, June Ock. 1988. «The impact of Confucianism on interpersonal relationships and communication patterns in East Asia.» *Communication Monographs* 55 (4): 374-388. Doi:10.1080/0393637758809376178.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 287-312
https://doi.org/10.18543/djhr.3197 • http://djhr.revistas.deusto.es/

Impacto de la inteligencia artificial en los derechos de los interesados: una perspectiva práctica

Impact of artificial intelligence on data subjects' rights: a practical overview

María Luisa González Tapia
Ramón y Cajal Abogados. España
mlgonzalez@ramoncajal.com
ORCiD: https://orcid.org/0009-0007-5281-5219

https://doi.org/10.18543/djhr.3198

Fecha de recepción: 30.05.2024
Fecha de aceptación: 25.10.2024
Fecha de publicación en línea: diciembre de 2024

Cómo citar / Citation: González Tapia, María Luisa. 2024. «Impacto de la inteligencia artificial en los derechos de los interesados: una perspectiva práctica». *Deusto Journal of Human Rights*, n. 14: 313-339. https://doi.org/10.18543/djhr.3198

Sumario: 1. Los derechos de los interesados. 1.1. Los derechos de los interesados, una garantía de control de los datos personales. 1.2. Aspectos prácticos del ejercicio de los derechos del interesado. Evolución desde la Directiva 95/46/CE al Reglamento General de Protección de Datos. 2. La IA como herramienta para el tratamiento de datos personales. 3. Peculiaridades de la atención del ejercicio de derechos en tratamientos que incluyen IA. 3.1. Rediseño de las políticas y procedimientos de atención de derechos internos. 3.2. Principales problemas generados por la utilización de sistemas y modelos de IA en el tratamiento de datos. 4. Decisiones automatizadas: el olvidado artículo 22 del RGPD. 4.1. Contenido del derecho reconocido en el artículo 22 del RGPD. 4.2. La dificultad de determinar cuándo nos encontramos ante una decisión automatizada. 4.3. Garantías que deben adoptarse si se supera la prohibición del artículo 22 del RGPD. Conclusiones. Bibliografía.

Resumen: Los denominados derechos del interesado, que aparecen regulados en el Capítulo III del Reglamento (UE) 2016/679, constituyen una de las principales herramientas puestas a disposición de los individuos para conseguir el control efectivo sobre sus datos personales. Entre dichos derechos figuran los de acceso, rectificación, supresión, oposición, portabilidad y limitación del tratamiento, además del poco ejercitado hasta la fecha derecho a no ser objeto de decisiones automatizadas. Como norma general, estos

derechos son directamente exigibles frente al responsable del tratamiento, que debe dar una respuesta formal dentro un plazo definido, incluso en aquellos supuestos en que corresponda la denegación de la solicitud. Este artículo analiza los cambios que la sucesiva implantación de sistemas de inteligencia artificial puede provocar en las peticiones de los afectados y los problemas prácticos a los que, previsiblemente, se enfrentarán los responsables del tratamiento que deban darles respuesta. En particular, resalta la importancia de realizar una adecuada interpretación y aplicación del artículo 22 del citado Reglamento (UE) 2016/679, relativo a decisiones automatizadas.

Palabras clave: Inteligencia artificial, protección de datos, derechos de los interesados, obligaciones del responsable del tratamiento, decisiones automatizadas.

Abstract: The so-called data subjects 'rights, regulated in Chapter III of Regulation (EU) 2016/679, are one of the most relevant guarantees available to individuals to gain effective control over their personal data. These rights include the rights of access, rectification, erasure, objection, portability and restriction of processing, in addition to the hitherto little-exercised right not to be subject to automated decisions. As a general rule, these rights are directly enforceable against the controller, who must provide a formal response within a defined period of time, even in those cases where a refusal of the request is appropriate. This article analyses the changes that the successive implementation of artificial intelligence systems may bring about in the requests of data subjects and the practical problems that data controllers who have to respond to them will probably face. In particular, it highlights the importance of a proper interpretation and application of Article 22 of Regulation (EU) 2016/679 on automated decisions.

Keywords: Artificial intelligence, data protection, data subjects' rights, data controller's obligations, automated decisions.

1. **Los derechos de los interesados**

1.1. *Los derechos de los interesados, una garantía de control de los datos personales*

El Capítulo III del 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 (en adelante, RGPD o Reglamento General de Protección de Datos) está dedicado a lo que se denominan "Derechos de los Interesados". Se inicia con el artículo 12 que recoge una serie de normas generales aplicables a todos ellos. En los preceptos siguientes, se regulan uno a uno los derechos propiamente dichos:

- Derecho a recibir información sobre el tratamiento (artículos 13 y 14).
- Derecho de acceso del interesado (artículo 15).
- Derecho de rectificación (artículos 16 y 19).
- Derecho de supresión o "derecho al olvido" (artículos 17 y 19).
- Derecho a la limitación del tratamiento (artículos 18 y 19).
- Derecho a la portabilidad de datos (artículo 20).
- Derecho de oposición (artículo 21).
- Derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles (artículo 22).

A excepción del primero y el último, los derechos del interesado tienen en común dos características básicas:

- Ser facultades del titular de los datos directamente exigibles frente al responsable del tratamiento, sin necesidad de ejercer acciones legales o presentar reclamaciones ante las autoridades de control, y
- obligar al responsable del tratamiento a una respuesta activa y formal (incluso cuando corresponda denegar la petición por forma abusiva o improcedente).

Podemos afirmar que se trata de derechos instrumentales que permiten un control real e inmediato del individuo sobre la información personal que le concierne. Constituyen, en definitiva, la materialización práctica del derecho fundamental a la protección de datos[1].

[1] Como señala Herrán (2002: 246), "el derecho a la autodeterminación informativa -como principio que otorga a la persona la posibilidad de determinar el nivel de

En este sentido, resulta relevante recordar que las Sentencias del Tribunal Constitucional 290/2000[2] y 292/2000[3], a través de la cuales comienza a definirse en nuestro ordenamiento jurídico el derecho fundamental a la protección de datos como un derecho diferenciado del derecho a la intimidad (Murillo de la Cueva 2007), resaltaron como característica del por entonces nuevo derecho las garantías de control que atribuye a las personas. Así, la Sentencia 290/2000, en su fundamento jurídico séptimo, señaló lo siguiente sobre el derecho fundamental a la protección de datos:

> (…) garantiza a la persona un poder de control y disposición sobre sus datos personales. Pues confiere a su titular un haz de facultades que son elementos esenciales del derecho fundamental a la protección de los datos personales, integrado por los derechos que corresponden al afectado a consentir la recogida y el uso de sus datos personales y a conocer los mismos. Y para hacer efectivo ese contenido, el derecho a ser informado de quién posee sus datos personales y con qué finalidad, así como el derecho a oponerse a esa posesión y uso exigiendo a quien corresponda que ponga fin a la posesión y empleo de tales datos.
>
> En suma, el derecho fundamental comprende un conjunto de derechos que el ciudadano puede ejercer frente a quienes sean titulares, públicos o privados, de ficheros de datos personales,

protección de los datos a ella referentes- encuentra su fundamento y su esencia en el reconocimiento de los derechos individuales con que la legislación otorga tutela a los interesados en la protección de datos personales. Los principios generales de protección de datos orientan y configuran la licitud del tratamiento de los datos personales, estableciendo los criterios elementales a seguir en el mismo. Las garantías individuales en la protección de datos constituyen instrumentos jurídicos al alcance de los interesados en defensa de sus derechos y libertades más esenciales".

[2] Sentencia del Tribunal Constitucional 290/2000, de 30 de noviembre (BOE n. 4, de 4 de enero de 2001). La sentencia resuelve los recursos de inconstitucionalidad acumulados n. 201/93, 219/93, 226/93 y 236/93, que fueron interpuestos, respectivamente, por el Consejo Ejecutivo de la Generalidad de Cataluña, el Defensor del Pueblo, el Parlamento de Cataluña y por D. Federico Trillo, Comisionado por 56 Diputados del Grupo Parlamentario Popular, contra los arts. 6.2, 19.1, 20.3, 22.1 y 2.1, 24, 31, 39.1 y 2, 40.1 y 2, y Disposición final tercera de la Ley Orgánica 5/1992, de 29 de octubre, de regulación del tratamiento automatizado de los datos de carácter personal (en adelante, LORTAD). Disponible en https://hj.tribunalconstitucional.es/es-ES/Resolucion/Show/4274

[3] Sentencia del Tribunal Constitucional 292/2000, de 30 de noviembre (BOE n. 4, de 4 de enero de 2001). La sentencia resuelve el recurso de inconstitucionalidad n. 1463-2000, interpuesto por el Defensor del Pueblo, contra los arts. 21.1 y 24.1 y 2 de la LORTAD. Disponible en: https://hj.tribunalconstitucional.es/es-ES/Resolucion/Show/4276

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

> partiendo del conocimiento de tales ficheros y de su contenido, uso y destino, por el registro de los mismos. De suerte que es sobre dichos ficheros donde han de proyectarse, en última instancia, las medidas destinadas a la salvaguardia del derecho fundamental aquí considerado por parte de las Administraciones Públicas competentes.

Teniendo en cuenta lo anterior, no parece extraño que, desde los primeros textos internacionales sobre protección de datos, se hayan reconocido ciertas facultades de control al interesado sobre el tratamiento que efectúa una determinada entidad o Administración Pública, si bien su formulación no ha sido siempre la misma. Han experimentado una evolución en su contenido, en cierta medida acorde a los nuevos usos de los datos personales y a las tecnologías o medios empleados por los responsables[4].

1.2. *Aspectos prácticos del ejercicio de los derechos del interesado. Evolución desde la Directiva 95/46/CE al Reglamento General de Protección de Datos*

El antecedente inmediato del actual marco normativo, la Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (Directiva 95/46/CE), regulaba, en sus artículos 12 a 15, un catálogo de derechos de los interesados dividido en dos bloques:

— Cuatro derechos activos: acceso, rectificación, supresión y oposición.
— Dos garantías adicionales que, en principio, deberían ser desplegadas por el responsable del tratamiento sin necesidad

[4] En las Directrices relativas a la protección de la intimidad y de la circulación transfronteriza de datos personales de la Organización para la Cooperación y el Desarrollo Económico de 1980 y, sobre todo, en el Convenio para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal, hecho en Estrasburgo el 28 de enero de 1981 (Convenio 108), aparecen ya unos derechos activos del titular de los datos que suponen un control práctico del tratamiento en los términos a los que nos hemos referido en los párrafos anteriores. En el Convenio 108, se denominan "garantías complementarias para las personas concernidas". Además, del acceso, se reconoce la facultad de rectificar o borrar aquellas informaciones que sobre su persona están siendo tratadas, con posibilidad de recurrir (no se aclara si a la jurisdicción ordinaria o a una autoridad independiente) en caso de que su petición no sea atendida.

de una petición previa del titular de los datos: el derecho de información y el derecho a no verse sometidas a una decisión con efectos jurídicos sobre ellas o que les afecte de manera significativa.

Durante los más de 20 años de aplicación de la citada norma y de sus leyes de transposición nacional, hablar de los derechos de los interesados significaba referirse a las facultades activas del primer bloque, que de manera muy resumida habilitaban al afectado para:

— solicitar que se le entregue una copia de los datos personales o se le informe de qué datos se están tratando;
— modificar datos incorrectos o desactualizados;
— pedir la supresión de los recogidos y almacenados por un responsable, y
— oponerse a la realización de determinados tratamientos, fundamentalmente, los relacionados con actividades comerciales.

El derecho de información se ha venido considerando más bien un deber del responsable del tratamiento, y no será objeto de análisis en este trabajo. Por su parte, el derecho a no verse sometido a determinadas decisiones automatizadas no ha tenido relevancia práctica hasta el momento. Como veremos, parece pensado para hacer frente a los retos actuales y, probablemente, se definirán mejor sus límites en los próximos años.

Tal y como los configuraba la Directiva 95/46/CE, los derechos activos influían en el tratamiento llevado a cabo por el responsable, que debía tomar determinadas medidas en función de lo solicitado por el titular de los datos como, por ejemplo, localizar la información referida a la persona que efectúa la petición en sus sistemas, determinar si procede suprimir o por el contrario está obligado a conservarlo, etc. En definitiva, el responsable se veía obligado a llevar a cabo una gestión de las peticiones recibidas, generalmente a través de procedimientos internos, que incluyen aspectos como:

— Establecer canales adecuados para la recepción de las peticiones.
— Formar al personal.
— Preparar modelos de respuesta.
— Acordar criterios para identificar a los solicitantes.
— Detallar supuestos de denegación de derechos.
— Desarrollar protocolos para la localización de los datos personales en los sistemas y archivos.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

La experiencia práctica consolidada con la Directiva 95/46/CE sirvió como base para que el Reglamento General de Protección de Datos reforzara y ampliara los derechos activos de los interesados, confiriéndoles mayor importancia, pero manteniendo el contenido esencial de todos ellos.

En primer lugar, como se ha indicado, el RGPD introdujo en su artículo 12 una serie de normas operativas para unificar criterios y aclarar dudas que se habían venido planteando en la resolución de solicitudes de afectados. Podemos resumir lo establecido en el referido precepto como sigue:

— Los derechos del interesado son gratuitos. No obstante, cuando las solicitudes sean infundadas o excesivas el responsable podrá cobrar un canon razonable para su ejercicio o negarse a atenderlo (artículo 12.5 del RGPD).
— Se dispone de un plazo de un mes para su atención, ampliable en dos meses adicionales por motivos justificados que han de comunicarse al afectado (art. 12.3 del RGPD).
— Deben responderse formalmente en todo caso, incluso cuando se deniegan. En este caso, se debe informar de las razones de su no actuación y de la posibilidad de presentar una reclamación ante una autoridad de control y de ejercitar acciones judiciales (artículo 12.4 del RGPD).
— Como norma general, el afectado no debe adjuntar copia de su DNI para ejercer el derecho. No obstante, el responsable del tratamiento que tenga dudas razonables sobre la identidad de la persona que efectúa la solicitud podrá pedir que se le facilite la información adicional necesaria (art. 12.6 RGPD).

En segundo lugar, el RGPD matizó el contenido de los derechos previstos en la Directiva 96/46/CE, también con la intención de armonizar criterios y positivizar prácticas implantadas.

En tercer y último lugar, se añadieron nuevos derechos que pretendían asegurar un mayor control del individuo sobre sus datos personales: el derecho a la portabilidad y el derecho a la limitación del tratamiento.

Con todo ello, se esperaba, además de conseguir la uniformidad a la que nos hemos referido en todos los Estados miembros, dar un impulso mayor y enfatizar el papel de los derechos del interesado como instrumentos garantes del correcto tratamiento de los datos.

En algunos Estados miembros, la regulación incluida en el RGPD implicó un cambio sustancial en la atención del ejercicio de derechos. En España, la Ley Orgánica 15/1999, de 13 de diciembre, de Protección

de Datos de Carácter Personal (LOPD) contaba con un reglamento de desarrollo, el Real Decreto 1720/2007, de 21 de diciembre, que estableció una completísima regulación de los denominados popularmente "Derechos ARCO" (acrónimo de acceso, rectificación, cancelación y oposición). Dicho Real Decreto había clarificado supuestos de denegación de derechos, establecido plazos para su atención, y detallado los requisitos formales para su ejercicio por parte del titular de los datos (entre los que se incluía la presentación del DNI o documento equivalente, que como sabemos ya no es un requerimiento obligatorio y cuya su solicitud puede dar lugar al incumplimiento del principio de minimización). Adicionalmente, se implantó un procedimiento específico en nuestra autoridad de control para tutelar los Derechos ARCO, denominado precisamente procedimiento de tutela de derechos.

Por ello, en nuestro país el Reglamento General de Protección de Datos no conllevó un cambio brusco en la gestión de las solicitudes de derechos de los afectados. La importancia que esta gestión ha adquirido en los últimos seis años se debe, probablemente, a diversos factores. Por una parte, ha aumentado el grado de información y concienciación entre los afectados de los derechos que les reconoce la normativa de protección de datos. Por otra, los tratamientos que se realizan son cada vez más complejos, y esto conlleva mayores molestias para los afectados y, también, mayores temores (por poner algunos ejemplos, es fácil que un tercero nos grabe o nos saque una fotografía y lo cuelgue en Internet, recibimos publicidad muy dirigida con cada compra o visita que efectuamos on-line, nos puede llegar a sorprender la exactitud con la que se predicen nuestros gustos en las plataformas de contenidos, etc.).

Una actividad que parecía residual en el contexto del cumplimiento de las obligaciones en materia de protección de datos para las entidades responsables del tratamiento ha llegado a consumir una parte importante del tiempo de los profesionales de la privacidad, especialmente, por su carácter casuístico. Los interesados que ejercen los derechos, como es lógico, no están obligados a conocer la normativa y a entenderlos. Sus peticiones no encajan siempre exactamente en una de las facultades reguladas.

En definitiva, al hablar de derechos de los interesados tratamos una materia más compleja y cambiante de lo que se venía pensando hasta la fecha, como se pone de manifiesto al constatar que las primeras previsiones del legislador europeo sobre la aplicación práctica del Título III del RGPD no se han cumplido. Así, por ejemplo, el nuevo derecho a la portabilidad de datos que se presentaba como un avance importante

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

para los afectados, hasta ahora, no ha supuesto ningún beneficio tangible para los afectados con respecto al marco normativo anterior[5].

Son probablemente, los derechos "clásicos" de acceso[6] y, en menor medida, de oposición y supresión (rebautizado como "derecho al olvido") los que siguen planteando los mayores problemas a la hora de atender resoluciones de los afectados. En el caso del derecho de oposición, su relevancia y complejidad va unida a la utilización del interés legítimo como base legal que habilita determinados tratamientos.

Con la adopción de forma masiva por parte de las empresas de sistemas y modelos de Inteligencia Artificial (en adelante, IA) se empieza a intuir que se producirá un cambio muy significativo en el ejercicio de los derechos de los interesados.

Desde la perspectiva de la normativa de protección[7], como ha señalado la Agencia Española de Protección de Datos (en adelante, AEPD), los sistemas y modelos de IA son un mero instrumento en el tratamiento de datos personales. No obstante, se trata de una herramienta de enorme relevancia para la protección de datos personales, tanto por el uso masivo que puede llevar a efectuarse de ellos durante el proceso de generación y entrenamiento de sistemas y modelos de IA, como en el resultado mismo obtenido de su aplicación.

---

[5] Uno de los primeros documentos elaborados por el Grupo de Trabajo del Artículo 29 (2017) tras la aprobación del Reglamento General de Protección de Datos fue el relativo al derecho de portabilidad, donde expone lo siguiente: "Las personas que hacían uso de su derecho de acceso en virtud de la Directiva sobre protección de datos 95/46/CE, se veían limitadas por el formato elegido por el responsable del tratamiento para proporcionar la información solicitada. El nuevo derecho a la portabilidad de los datos tiene por objeto facultar a los interesados con respecto a sus propios datos personales, ya que mejora su capacidad de trasladar, copiar o transmitir datos personales fácilmente de un entorno informático a otro (ya sea a sus propios sistemas, a los sistemas de terceros de confianza o a los de otros responsables del tratamiento). Al afirmar los derechos personales de los individuos y el control sobre los datos personales que les conciernen, la portabilidad de los datos representa también una oportunidad para «reequilibrar» la relación entre los interesados y los responsables del tratamiento".

[6] Así lo ha reconocido el Comité Europeo de Protección de Datos (2023) dedicándole un documento.

[7] Según indica la Agencia Española de Protección de Datos (2023), "un sistema de Inteligencia Artificial (IA), o varios sistemas de IA, podría ser un medio seleccionado por un responsable para implementar operaciones de datos personales en un tratamiento. Es importante entender que la finalidad última de un tratamiento es diferente de los medios seleccionados para implementarlo. Con relación a esto, el responsable será quien determine si los resultados de un sistema de IA implicarían una decisión automática o determinará que se incluya una supervisión humana que tome la decisión final. Por lo tanto, las decisiones automatizadas no están en la naturaleza del sistema de IA, sino que son una opción elegida por el responsable".

2. La IA como herramienta para el tratamiento de datos personales

Coloquialmente, al hablar de IA nos referimos a programas informáticos que simulan la mente humana, y en particular, que tienen capacidad de aprendizaje y adaptación al entorno: pueden elaborar un texto complejo (por ejemplo, un relato) sobre unas premisas definidas, mantener una conversación en un servicio de atención al cliente, dar respuesta a preguntas en tiempo real, o generar imágenes que plasmen una descripción dada por una persona.

La IA no es una tecnología nueva pero su uso se ha generalizado de manera rápida en los últimos años. En 2017, la Comisión Europea estimaba que podría generar entre 6.5 y 12 billones de euros para el 2025 (Comisión Europea 2017). Al año siguiente, en el 2018, señalaba que

> al igual que hicieran la máquina de vapor o la electricidad en épocas anteriores, la IA está transformando nuestro mundo, nuestra sociedad y nuestra industria. El crecimiento de la capacidad informática y la disponibilidad de datos, así como los avances en los algoritmos, han convertido la IA en una de las tecnologías más estratégicas del siglo XXI (Comisión Europea 2018).

Como resulta lógico, afirmaciones como las anteriores se han visto acompañadas de la elaboración del marco regulatorio de la IA dentro en la Unión Europea. A finales de 2022, la Comisión Europea presentó dos propuestas legislativas en este sentido:

— Una propuesta de revisión de la Directiva 85/374/CE del Consejo, de 25 de julio de 1985, relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados Miembros en materia de responsabilidad por los daños causados por productos defectuosos.
— Una propuesta de norma específicamente dirigida a la IA, que acabó convirtiéndose en el Reglamento de Inteligencia Artificial aprobado por el Parlamento Europeo el pasado 13 de marzo (en adelante, RIA), y que se calificó en ese momento de "ley histórica" (Parlamento Europeo 2024).

El RIA tiene un enfoque de seguridad de producto, del que se regula su fabricación, puesta en mercado, distribución y uso. El producto en cuestión son los sistemas y modelos de IA. Los primeros (sistemas), se definen en su artículo 3 como aquellos basados

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

> en una máquina diseñado para funcionar con distintos niveles de autonomía, 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 información de salida, como predicciones, contenidos, recomendaciones o decisiones, que puede influir en entornos físicos o virtuales (DOUE 2024).

Frente a este concepto, los modelos de IA son códigos o funciones, más simplificados que los sistemas que no incluyen una variedad de componentes.

La definición legal arriba reproducida, de la que probablemente se publicarán documentos aclaratorios, incluye las mismas características básicas que forman parte de la idea que se asocia comúnmente entre los no expertos al concepto de IA: autonomía, adaptación y obtención de resultados sofisticados. Nos interesa destacar esto último, la capacidad de generar información de salida inferida de información de entrada, puesto que incide en la condición de herramienta o instrumento de los sistemas y modelos de IA, y en el valor que los datos, ya sean personales o no, tienen en su funcionamiento.

Resulta fácil imaginar que las mejoras que promete la utilización de IA en distintos campos no están exentas de riegos para los derechos y libertades de los ciudadanos. En materia de protección de datos, se debe resaltar que la IA ampliará las posibilidades de generación de contenidos y de realización de predicciones de todo tipo[8]. Tales predicciones pueden consistir en decisiones automatizadas que, por ejemplo, evalúen a los individuos considerándolos aptos o no aptos

---

[8] Según se expone en el documento elaborado conjuntamente por el Comité Europeo de Protección de Datos y el Supervisor Europeo de Protección de Datos (2021) sobre la nueva norma de inteligencia artificial*:*

> "Los datos (personales y no personales) en la IA son, en muchos casos, la premisa clave de las decisiones autónomas, lo que inevitablemente afectará a la vida de las personas a distintos niveles.
>
> Asignar a las máquinas la tarea de decidir a partir de datos creará riesgos para los derechos y libertades de las personas, afectará a su vida privada y podría perjudicar a algunos grupos o incluso a las sociedades en su conjunto. El CEPD y el SEPD subrayan que el derecho a la vida privada y a la protección de los datos personales, que contradicen con la asunción de la autonomía de decisión de las máquinas que subyace al concepto de IA, es un pilar de los valores de la UE reconocidos en la Declaración Universal de Derechos Humanos (artículo 12), el Convenio Europeo de Derechos Humanos (artículo 8) y la Carta de los Derechos Fundamentales de la UE (en lo sucesivo, «la Carta») (artículos 7 y 8). Conciliar la perspectiva de crecimiento que ofrecen las aplicaciones de IA y la centralidad y primacía de los seres humanos frente a las máquinas es un objetivo muy ambicioso, pero necesario".

para un puesto de trabajo o para obtener un determinado beneficio, o los categoricen en determinados grupos o perfiles en función de parámetros que pueden resultar erróneos o discriminatorios.

El mayor peligro consiste en que, a medida que se confía en un sistema o modelo de IA para efectuar una tarea como las indicadas, se pierde la capacidad de análisis detallado del supuesto concreto y de la relación de causalidad entre la información de la que se parte y el resultado obtenido.

El cumplimiento de las obligaciones contenidas en el RIA no exime del cumplimiento del RGPD en todas las fases en las que se utilicen datos personales. Tal y como señala su Considerando (9),

> las normas armonizadas que se establecen en el presente Reglamento deben aplicarse en todos los sectores y, en consonancia con el nuevo marco legislativo, deben entenderse sin perjuicio del Derecho vigente de la Unión, en particular en materia de protección de datos, protección de los consumidores, derechos fundamentales, empleo, protección de los trabajadores y seguridad de los productos, al que complementa el presente Reglamento. En consecuencia, permanecen inalterados y siguen siendo plenamente aplicables todos los derechos y vías de recurso que el citado Derecho de la Unión otorga a los consumidores y demás personas que puedan verse afectados negativamente por los sistemas de IA (DOUE 2024).

Por tal motivo, resulta fácil entender que la introducción de herramientas de IA afectará a distintos aspectos de la protección de datos, y en particular, a la forma en la que se atienden los derechos de los interesados que parecen en este contexto más necesarios que nunca para asegurar un control efectivo sobre los datos personales.

3. **Peculiaridades de la atención del ejercicio de derechos en tratamientos que incluyen IA**

3.1. *Rediseño de las políticas y procedimientos de atención de derechos internos*

Comenzaremos indicando que la aparición de componentes de IA en un tratamiento de datos no conlleva, en principio, ninguna excepción en lo que se refiere al ejercicio de los derechos del afectado. Así lo señala expresamente la AEPD en uno de los documentos que ha dedicado al análisis de tratamientos que incluyen IA:

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

> Los responsables que hagan uso de soluciones de IA para tratar datos personales, elaborar perfiles o tomar decisiones automatizadas, han de ser conscientes de que los interesados tienen derechos en el ámbito de la protección de datos que deben ser atendidos.
>
> Por lo tanto, durante la fase de concepción del tratamiento, los responsables han de ser conscientes de que tienen que establecer mecanismos y procedimientos adecuados para poder atender las solicitudes que reciban, y que dichos mecanismos deberán estar adecuadamente dimensionados para la escala del tratamiento que están efectuando (Agencia Española de Protección de Datos 2020).

El responsable del tratamiento, que seguramente ya dispone de procedimientos de atención al ejercicio de derechos del afectado desde antes de la aprobación del RGPD, deberá plantearse ahora nuevos supuestos.

Hemos reseñado que la gestión de derechos se ha convertido en un aspecto complejo en un gran número de organizaciones y, en este contexto, la IA supondrá un mayor nivel de dificultad. Por tanto, resulta recomendable (i) llevar a cabo un estudio detallado de las posibles peticiones de los afectados en tratamientos y (ii) documentar criterios fundamentados jurídicamente tanto en relación con la forma en la que harán efectivas solicitudes como con los supuestos en los que se denegarán.

Sin este análisis no es posible adoptar mecanismos realistas y proporcionales para garantizar el ejercicio de derechos adaptándose al nivel de riesgo de cada tratamiento. Además, se ha de tener en cuenta que la utilización de sistemas y modelos de IA en las operaciones de tratamiento tendrá que ir acompañada de la realización o revisión de los análisis de riesgos y de las evaluaciones de impacto (que probablemente se deberán efectuar por lo novedoso de la tecnología, los volúmenes de datos utilizados y el carácter potencialmente intrusivo). Estos documentos internos deberían contemplar una descripción de cómo se van a atender los derechos en cada caso.

3.2. *Principales problemas generados por la utilización de sistemas y modelos de IA en el tratamiento de datos*

Entendemos que los cinco principales problemas que surgirán en el análisis y rediseño de los procedimientos de atención al ejercicio de derechos son los siguientes:

3.2.1. Dificultad en la localización de los datos del interesado

Normalmente, esta dificultad proviene del volumen de datos tratados o su dispersión en distintos sistemas que no han sido adecuadamente inventariados. No obstante, en algunos casos, también se deberá al hecho de no conocer si se están tratando o no datos personales en una etapa concreta del proceso[9].

Por ejemplo, durante la fase de entrenamiento de sistemas o modelos, donde se maneja una gran cantidad de información, los datos personales de base pueden someterse a procesos de anonimización (eliminación de cualquier posibilidad de vinculación con el individuo al que se refieren) o seudonimización.

Los procesos de anonimización constituyen tratamientos de datos en sí. Al menos en su inicio, partimos de datos personales en relación a los cuales los interesados pueden ejercer sus derechos. El ejercicio de derechos en estos supuestos no resulta baladí, sobre todo cuando se solicita la oposición a tratamientos basados en el interés legítimo y facilitar tal oposición constituye una garantía esencial del afectado. Adicionalmente, la anonimización no siempre es efectiva, y podría existir la posibilidad de reidentificación (y, por tanto, tratamiento de datos personales) en un futuro, bien por errores no previstos o por la mera evolución técnica[10].

Por lo que se refiere a la seudonimización, resulta necesario destacar que la dificultad en la identificación de un afectado no implica que no se estén tratando datos personales y, consiguientemente, no excluye la atención del ejercicio de derechos. En este sentido, a veces

[9] A modo de ejemplo, señalaremos lo indicado por la Agencia Española de Protección de Datos (2020): "Puede haber tratamientos que incluyen componentes de IA que manejan datos de personas físicas, como en un modelo de perfilado de marketing o electoral, o puede haber tratamientos en los que no aparezcan datos de carácter personal, como podría suceder en un modelo de predicción meteorológico que recoge datos de estaciones geográficamente distribuidas. Un tratamiento que tome decisiones automatizadas usando la inteligencia artificial puede afectar a personas físicas, como por ejemplo un sistema de autenticación de usuarios, o puede no afectar a personas, como un sistema de control industrial. En el que caso de que se tomen decisiones que afectan a las personas, estas decisiones pueden ser relativas a la interacción de la persona en su contexto social, como el acceso a un contrato o servicio, o relativas a la personalización de dicho servicio, como podría ser la personalización en los mandos de un coche o la programación de un televisor".

[10] La Agencia Española de Protección de Datos (2016) y Personal Data Protection Commission Singapore (2022) han publicado en su página web distintos documentos orientativos sobre cómo realizar procesos de anonimización correctos, donde se resaltan los riesgos de reidentificación producidos por la evolución de las tecnologías.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

resulta de utilidad solicitar información adicional al afectado que ejerce un derecho a los efectos de obtener nuevos parámetros de búsqueda.

En otras ocasiones, tal y como pone de manifiesto la autoridad de protección de datos de Reino Unido, la *Information Commissioner's Office* (ICO) (s.f.), aunque el responsable debe atender la solicitud de todos los interesados, pueden darse supuestos en los que la petición resulte excesiva o desproporcionada. Por tanto, sería posible denegarla. La lógica nos dice que no podemos olvidar que la seudonimización es una garantía para el propio afectado, ya que reduce la información personal que se somete a tratamiento, y su aplicación no debería penalizar al responsable a la hora de cumplir con otras obligaciones como la atención del ejercicio de derechos haciéndola más complicada. Dado que siempre corresponderá la carga de la prueba al responsable, es aconsejable documentar y justificar los escenarios de denegación de peticiones.

3.2.2. Imposibilidad de eliminación o modificación de los datos personales

Algunas soluciones tecnológicas no permiten suprimir definitivamente los datos o ejecutar cambios en los mismos y, por tanto, imposibilitan en sí mismas la correcta atención de los derechos de los afectados.

El derecho de supresión resulta, por otro lado, especialmente complicado de atender por las dudas que, inevitablemente, surgen sobre (i) los plazos de conservación aplicables y (ii) la procedencia de aplicar la figura del bloqueo recogida en el artículo 32 de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD).

Según este precepto, cuando un afectado ejerce el derecho de supresión, los datos no se deben eliminar sino conservarse aplicando medidas técnicas y organizativas que impidan cualquier tipo de utilización, incluyendo el mero acceso, excepto para la puesta a disposición de los datos a los jueces y tribunales, el Ministerio Fiscal o las Administraciones Públicas competentes, en particular de las autoridades de protección de datos.

El análisis de riesgos y la evaluación de impacto realizados en las primeras fases del tratamiento han de contemplar aspectos que luego simplificarán la atención del ejercicio de derechos como plazos de conservación, medidas de bloqueo de datos, revisión de mecanismos técnicos para garantizar la eliminación o cancelación de la información, o existencia de canales apropiados para el ejercicio de derechos.

También deberían reseñar si, en el momento del diseño de la infraestructura o sistema de tratamiento, se han incluido medidas que permitan suprimir o bloquear datos en caso de ser necesario.

Entendemos que, en última instancia, se trata de una obligación que corresponde al proveedor o fabricante, pero el responsable del tratamiento que actúe, por utilizar la terminología del RIA, como responsable del despliegue ha de aplicar medidas de diligencia en la selección del producto que utiliza.

3.2.3. Dependencia de terceros proveedores que actúan como encargados del tratamiento

En la misma línea expuesta en el punto anterior, cuando parte de los datos personales se alojan en los sistemas del encargado, siendo además éste quien trabaja sobre los mismos, se han de establecer mecanismos de comunicación que permitan al responsable verificar que los derechos se ejecutan correctamente.

En este sentido, la adopción de medidas previas a la contratación para verificar las garantías que ofrece el proveedor para una correcta atención del ejercicio de derechos podría resultar cada vez más relevante, así como la inclusión de cláusulas específicas en los contratos de encargo del tratamiento.

En todo caso, la utilización de encargados del tratamiento tampoco exime al responsable del cumplimiento de sus obligaciones de atención de los derechos del afectado. Recordemos que los proveedores de soluciones tecnológicas también quedan obligados por los principios de privacidad desde el diseño y por defecto y deberían desarrollar sus productos incluyendo garantías que permitan una correcta atención de los derechos.

3.2.4. Asignación de recursos a la gestión interna de los derechos

Como consecuencia de los puntos anteriores, es posible que en algunas organizaciones la atención de derechos requiera que se involucren perfiles distintos a los habituales, especialmente con un componente más técnico que permita comprender el funcionamiento de los sistemas y modelos de IA.

La formación y la comprensión del funcionamiento de los sistemas y modelos de IA resultará básica para poder dar una respuesta adecuada a las peticiones recibidas de los interesados.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

3.2.5. Ejercicios de derechos relacionados con el resultado de la aplicación de IA o información de salida

Siendo los anteriores puntos muy relevantes, el mayor impacto previsiblemente se producirá en el ejercicio de derechos en la fase de obtención de resultados. En esos casos, desde el punto de vista de protección de datos, cobra relevancia determinar si los elementos de IA sirven para obtener perfiles individualizados o tomar decisiones automatizadas. Ambas actividades se consideran generalmente intrusivas.

El art. 4.4 del RGPD determina que "elaboración de perfiles" consiste en

> toda forma de tratamiento automatizado de datos personales consistente en utilizar datos personales para evaluar determinados aspectos personales de una persona física, en particular para analizar o predecir aspectos relativos al rendimiento profesional, situación económica, salud, preferencias personales, intereses, fiabilidad, comportamiento, ubicación o movimientos de dicha persona física.

Esta definición implica que se cumplan tres requisitos acumulativos:

— es un tratamiento automatizado;
— emplea datos personales; y
— el objetivo es realizar una evaluación o juicio aplicable a un individuo concreto con una finalidad de seguimiento o de tratamiento también individualizada. Conocer el perfil genérico o tipo de cliente (por ejemplo, que el comprador típico de un producto es mujer profesional de 30 años) tras un análisis de las características de los compradores de una tienda on-line no implica la elaboración de un perfil, aunque para ello se hayan partido del tratamiento de datos personales de un número relevante de compradores.

El RGPD introduce posteriormente en su articulado otro concepto que no define: decisión basada únicamente en el tratamiento automatizado. Las decisiones automatizadas parecen tener un ámbito de aplicación distinto a la elaboración de perfiles, aunque pueden solaparse. Suponen la utilización medios tecnológicos para suplantar la capacidad humana en actividades tales como asignar, seleccionar o eliminar candidatos a puestos de trabajo, predecir el nivel de riesgo de impago de un préstamo, aplicar unos determinados parámetros para asignar turnos o proponer compras, etc. El perfilado podría ser un tipo

de decisión automatizada, pero las decisiones automatizadas no implican necesariamente elaborar perfiles. Por otro lado, una vez perfilado un individuo, se podrían tomar decisiones que le afecten de forma automatizada o no.

Los conceptos anteriores no son fáciles de entender y diferenciar, aunque se intuyen potencialmente peligrosos para los derechos del afectado. El artículo 22 del RGPD reconoce un derecho específico relacionado con ambas operaciones, bajo el epígrafe "decisiones individuales automatizadas, incluyendo la elaboración de perfiles". A este derecho, no se le ha prestado demasiada atención hasta la fecha (raramente se menciona en los textos informativos de privacidad). Su contenido y utilidad práctica se irá concretando en el futuro como veremos a continuación.

4. **Decisiones automatizadas: el olvidado artículo 22 del RGPD**

4.1. *Contenido del derecho reconocido en el artículo 22 del RGPD*

El Capítulo III del RGPD se cierra con el reconocimiento de un derecho de los interesados que se formula en el artículo 22.1 como sigue:

> Todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar.

Si bien parece introducido específicamente para responder a las necesidades que presenta el momento actual, en el que la IA parece que se convertirá en la herramienta básica de trabajo de muchas organizaciones, el origen de este derecho se remonta a la primera ley de protección de datos francesa de 1970. También nuestra primera ley orgánica de protección de datos, Ley Orgánica 5/1992, de 29 de octubre, conocida como LORTAD, reconocía entre los derechos de las personas el derecho a la impugnación de valoraciones basadas exclusivamente en datos automatizados. Por su parte, la Directiva 95/46/CE reguló un derecho en términos similares a los del citado artículo 22 del RGPD: el derecho a no verse sometido a determinadas decisiones automatizadas, dando lugar a transposiciones que diferían sustancialmente en los distintos Estados miembros. En la transposición al derecho nacional, algunos como Bélgica establecieron una

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

prohibición, mientras que otros, como Suecia, garantizaron un sistema de *opt-out* u oposición.

En definitiva, no nos encontramos ante una facultad o derecho totalmente nuevo. Sin embargo, ha sido poco utilizado en la práctica y no ha dado lugar ni a la imposición de número elevado de sanciones por las autoridades de control nacionales ni a resoluciones o sentencias que interpreten su contenido.

En 2017, previendo la relevancia que podría tener por el auge en la utilización de IA, el desaparecido Grupo de Trabajo del Artículo 29 (GT 29) (2018) le dedicó un documento específico. En su introducción, se apuntan los problemas y riesgos derivados del uso masivo de datos y la irrupción de la IA:

> La elaboración de perfiles y las decisiones automatizadas se utilizan en un número creciente de sectores, tanto privados como públicos.
>
> El sector bancario y financiero, la asistencia sanitaria, la fiscalidad, los seguros, la mercadotecnia y la publicidad son solo algunos ejemplos de los ámbitos en los que se lleva a cabo con más regularidad la elaboración de perfiles para contribuir al proceso de toma de decisiones.
>
> Los progresos tecnológicos y las posibilidades del análisis de macrodatos, la inteligencia artificial y el aprendizaje automático han facilitado la creación de perfiles y han automatizado las decisiones, y tienen el potencial de afectar de forma significativa a los derechos y libertades de las personas. (…)
>
> No obstante, la elaboración de perfiles y las decisiones automatizadas pueden plantear riesgos importantes para los derechos y libertades de las personas que requieren unas garantías adecuadas.
>
> Estos procesos pueden ser opacos. Puede que las personas no sean conscientes de que se está creando un perfil sobre ellas o que no entiendan lo que implica.
>
> La elaboración de perfiles puede perpetuar los estereotipos existentes y la segregación social. Asimismo, puede encasillar a una persona en una categoría específica y limitarla a las preferencias que se le sugieren. Esto puede socavar su libertad a la hora de elegir, por ejemplo, ciertos productos o servicios como libros, música o noticias. En algunos casos, la elaboración de perfiles puede llevar a predicciones inexactas. En otros, puede llevar a la denegación de servicios y bienes, y a una discriminación injustificada.

La anterior descripción podría servir para ejemplificar los efectos negativos de los sesgos o *bias* que pueden darse en sistemas de IA.

Por ello, el documento del GT 29 constituye una herramienta que ha cobrado gran utilidad: las recomendaciones que se establecen en el mismo son en la práctica un listado de controles críticos de cumplimiento para sistemas y modelos de IA (entre otros, cumplimiento de principio de licitud y transparencia, minimización de datos y respeto al principio de finalidad).

Centrándonos en la interpretación del derecho reconocido en el artículo 22, el GT 29 aclara algo que había sido objeto de discusión hasta ese momento por las diferencias en las transposiciones nacionales de la Directiva 95/46/CE que se han mencionado: el apartado 1 de este precepto debe interpretarse como una prohibición genérica de decisiones automatizadas que produzcan efectos que podemos definir como trascendentes en los afectados. No requiere una actividad del afectado (una solicitud formal) para que el artículo despliegue sus efectos.

Como señala el GT 29 esta

> interpretación refuerza la idea de que sea el interesado quien tenga el control sobre sus datos personales, lo cual se corresponde con los principios fundamentales del RGPD. Interpretar el artículo 22 como una prohibición en vez de como un derecho que debe invocarse significa que las personas están protegidas automáticamente frente a las posibles consecuencias que pueda tener este tipo de tratamiento.

Además, el GT 29 pone el derecho en relación con el Considerando 71 del RGPD, que determina:

> Sin embargo, se deben permitir las decisiones basadas en tal tratamiento, incluida la elaboración de perfiles, si lo autoriza expresamente el Derecho de la Unión o de los Estados miembros [...], o necesario para la conclusión o ejecución de un contrato [...], o en los casos en los que el interesado haya dado su consentimiento explícito.

La prohibición del artículo 22, en todo caso, no es absoluta, ya que tiene dos límites:

a) Por un lado, solo se aplica a decisiones totalmente automatizadas que, además, han de producir (i) efectos jurídicos, como la resolución de un contrato o la retirada de un beneficio social, o (ii) afectar significativamente de modo similar a las personas. En este último punto, deberíamos incluir

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

consecuencias relevantes para el afectado como su no selección para un puesto de trabajo e, incluso, determinadas acciones promocionales que tengan un carácter intrusivo. A modo de ejemplo, podemos señalar las campañas dirigidas a menores o colectivos vulnerables (personas susceptibles de contratar préstamos rápidos o jugar on-line).

A sensu contrario, si no existe una automatización total o los efectos de la decisión no son ni jurídico ni especialmente relevantes, podrá efectuarse el tratamiento.

b) Por otro lado, el apartado segundo del mismo artículo 22 señala tres excepciones que permiten llevar a cabo las decisiones automatizadas inicialmente prohibidas. Estas excepciones se dan en los siguientes supuestos:

— Cuando la decisión automatizada es necesaria para la celebración o la ejecución de un contrato entre el interesado y un responsable del tratamiento;
— Cuando está autorizada por el Derecho de la Unión o de los Estados miembros, que además deberá establecer medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado, o
— Cuando se basa en el consentimiento explícito de los interesados.

En definitiva, el legislador protege a los interesados con el reconocimiento de un derecho que se plasma en la existencia de una prohibición general, similar a la del artículo 9.1 del RGPD. Como en el caso de este último precepto, se han regulado excepciones que levantan dicha prohibición.

4.2. *La dificultad de determinar cuándo nos encontramos ante una decisión automatizada*

La principal tarea a la que se enfrentan los responsables del tratamiento a la hora de garantizar el derecho del artículo 22 es la de catalogar una decisión como completamente automatizada e identificar los efectos que puede producir en el afectado. El derecho no obliga a prever respuestas o a reaccionar cuando el afectado lo pide, sino más bien a justificar un tratamiento desde su origen, analizando si es lícito de acuerdo con el Reglamento General de Protección de Datos.

Como se ha explicado, no existe un amplio abanico de resoluciones que interpreten el artículo 22 que comentamos[11], que por otra parte es muy casuístico. En España, probablemente, la resolución más relevante de la AEPD que analiza el concepto de decisiones automatizadas se refiere a un asunto transfronterizo que se cerró con una multa de 550.000 euros a la entidad GLOVOAPP23, S.A. (en adelante, GLOVOAPP). En este caso, se investigó la utilización de una aplicación que asignaba turnos y pedidos a los denominados *riders* de forma automatizada y era empleada por las distintas filiales de la sociedad sancionada en varios Estados miembros[12]. De hecho, las actuaciones se inician en Italia, con una inspección Foodinho SRL. por parte del *Garante per la Protezione dei Dati Personali*.

Uno de los aspectos que se analiza en la resolución, no el único, es si la aplicación implica la toma de decisiones automatizadas. GLOVOAPP defiende que no se cumplen los requisitos para ello, en síntesis, porque los criterios de asignación de turnos estaban decididos o acordados con intervención humana, limitándose la aplicación a llevar a cabo un proceso automatizado consistente en tres fases: (i) Aplicar una tabla de franjas horarias acordada por las partes; (ii) ejecutar una decisión adoptada por las partes; (iii) dar acceso a franja horaria según un orden establecido previamente en función de las preferencias de las partes.

Sin embargo, en opinión de la AEPD, la decisión automatizada existe, ya

> que es el sistema el que adoptaba la decisión sobre en qué orden se permitía acceder a unos repartidores determinados para la reserva de una franja horaria concreta, independientemente de que era GLOVOAPP como responsable de tratamiento quien introducía los parámetros necesarios en el Sistema para que pudiera adoptar tal

[11] Una completa revisión de las resoluciones recientes de autoridades de control puede encontrase en Future of Privacy Forum (2022). Igualmente, resulta interesante la lectura de Privacy Internacional (2017).

[12] Nos referimos al procedimiento sancionador con referencia PS/00209/2022, disponible en https://www.aepd.es/documento/ps-00209-2022.pdf. Este procedimiento fue objeto de recurso de reposición, pudiéndose consultar su resolución en: https://www.aepd.es/documento/reposicion-ps-00209-2022.pdf

Se imputan a la entidad sancionada las siguientes infracciones:

— Infracción del artículo 13 del RGPD, por la que se le apercibe.
— Infracción de los artículos 25 (privacidad desde el diseño y por defecto) y 32 (aplicación de medidas de seguridad) del RGPD, por las que se le impone la multa de 550.000 €.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

> decisión. La decisión sobre el orden en que se permitía acceder a los repartidores a las franjas horarias era adoptada por la aplicación, sin intervención humana de ningún tipo. Únicamente se producía una intervención humana en aquellos casos en que los repartidores reclamaban, pero si no había reclamación, no había intervención humana que modificara dicha decisión ni supervisión alguna de tal decisión.

Más recientemente, hemos conocido la sentencia del Tribunal de Justicia de la Unión Europea (TJUE) del 7 de diciembre de 2023, que es la primera en abordar centralmente el artículo 22, con un enfoque amplio y garantista que extiende el concepto de decisión automatizada hasta la actividad de una empresa que facilita información de solvencia o probabilidad de impago a terceros que son los que toman realmente la decisión que afecta al individuo[13].

Previendo que se dará una interpretación amplia a los conceptos de decisión automatiza y decisión completamente automatizada, será importante demostrar que o bien los efectos en los interesados no tienen consecuencias relevantes, o bien que se cumplen las excepciones ya comentadas del artículo 22.2.

Por ello, las recomendaciones a la hora de revisar estos tratamientos son (i) documentar y justificar la licitud del tratamiento y (ii) aplicar garantías adicionales suficientes.

4.3. *Garantías que deben adoptarse si se supera la prohibición del artículo 22 del RGPD*

Cuando existan decisiones automatizadas excluidas de la prohibición del artículo 22 deberán tomarse determinadas garantías, que incluyen, además del cumplimiento del resto de las obligaciones del Reglamento General de Protección de Datos, las siguientes:

— No utilizar categorías especiales de datos, salvo que se cuente con el consentimiento explícito del afectado (artículo 9.2.a del RGPD) o el tratamiento sea necesario por razón de un interés público esencial (artículo 9.2.g del RGPD).

— Informar a los afectos de forma expresa y específica de la lógica aplicada para tomar la decisión, así como de las consecuencias

[13] Para un análisis detallado puede consultarse Cotino (2024).

previstas e importancia de la mismas (según se recoge en los artículos 13 y 14 del RGPD)[14].

— Atender los derechos activos y directamente ejercitables frente al responsable del tratamiento que se recogen en el apartado 3 del propio artículo 22: derecho a obtener intervención humana, a expresar su punto de vista y a impugnar la decisión. En principio, parecen tres facultades que pueden ejercerse de forma diferenciada y en cuya tramitación deberán aplicarse las normas generales ya citadas que se incluyen en el artículo 12 del RGPD. Se limita la posibilidad de utilizarlos a los supuestos en los que las decisiones automatizadas se hayan basado en la existencia de un contrato o en el consentimiento explícito del afectado. En este punto, conviene advertir que, si su solicitud se generaliza, significará que los responsables del tratamiento deberán arbitrar procedimientos de revisión de las decisiones adoptadas a petición de los titulares de los datos.
— La realización de una evaluación de impacto sobre la protección de datos, como en el caso de otros tratamientos que entrañen un riesgo sustancial para los derechos y libertades de los afectados.

Conclusiones

La utilización de sistemas y modelos de IA de forma generalizada en las organizaciones hará más complejo el cumplimiento de las obligaciones derivadas de la normativa de protección de datos. La atención del ejercicio de derechos, que ha adquirido relevancia en los últimos años por su carácter casuístico, se enfrenta a problemas nuevos como la dificultad de localizar los datos personales del afectado, la mayor dependencia de proveedores tecnológicos muy especializados, o el propio desconocimiento de si realmente se tratan datos personales en algunas fases como el entrenamiento de sistemas de IA.

El mayor reto para los profesionales de la privacidad a corto plazo consiste en recibir una formación adecuada que permita comprender los tratamientos que se están llevando a cabo y las funcionalidades de los componentes de IA que aparecen en los mismos. Aquellos que

[14] Por lo que se refiere al ámbito laboral, el Ministerio de Trabajo y Economía Social publicó en 2022 una guía en la que se explicaba el derecho de los afectados a la recibir lo que se denomina "información algorítmica" tomando como punto de partida el artículo 22 del RGPD (Ministerio de Trabajo y Economía Social 2022).

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 313-339
https://doi.org/10.18543/djhr.3198 • http://djhr.revistas.deusto.es/

desempeñen funciones de delegados de protección de datos tendrán que intervenir de forma activa en la evaluación de los riesgos asociados a la utilización de sistemas y modelos de IA, y quizás se coordinen sus funciones con responsables de IA.

Las autoridades de protección de datos están publicando diversos documentos y guías para ayudar a las organizaciones en el cumplimiento de sus obligaciones. Esperamos que también se aprueben documentos aclaratorios del RIA en los próximos meses. Probablemente, también en un corto plazo veremos resoluciones y sentencias que dan forma al concepto de decisión automatizada. En este sentido, resulta relevante entender qué significa "sin intervención humana", es decir cuál es el grado de participación o revisión de una persona física que se requiere para que una decisión no sea totalmente automatizada.

El artículo 22 del RGPD y las obligaciones de transparencia sobre la lógica aplicada para tomar una decisión que prevén los artículos 13 y 14 de la misma norma resultan claves para asegurar el control de los interesados sobre sus datos personales. También lo es la posibilidad de conocer los datos personales tratados durante un proceso en el que intervienen componentes de IA.

Desde el punto de vista del individuo, las garantías establecidas por el RGPD son un complemento necesario que refuerza la protección que les otorga el RIA frente a sistemas y modelos de inteligencia artificial con efectos discriminatorios o impactos negativos en otros derechos y libertades.

Bibliografía

Agencia Española de Protección de Datos. 2016. *Orientaciones y garantías en los procedimientos de anonimización de datos personales*. Acceso el 10 de octubre de 2024: https://www.aepd.es/guias/guia-orientaciones-procedimientos-anonimizacion.pdf

Agencia Española de Protección de Datos. 2020. *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción.* Acceso el 10 de octubre de 2024: https://www.aepd.es/documento/adecuacion-rgpd-ia.pdf

Agencia Española de Protección de Datos. 2023. «Inteligencia Artificial: Sistema vs. tratamiento, medios vs. Finalidad» (post de 10 de abril). Acceso el 10 de octubre de 2024: https://www.aepd.es/prensa-y-comunicacion/blog/inteligencia-artificial-sistema-vs-tratamiento-medio-vs-finalidad

Comité Europeo de Protección de Datos-Supervisor Europeo de Protección de Datos. 2021. *Dictamen conjunto 5/2021 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). Acceso el 10 de octubre de 2024: https://www.edpb.europa.eu/system/files/2021-10/edpb-edps_joint_opinion_ai_regulation_es.pdf

Comité Europeo de Protección de Datos. 2023. *Directrices 01/2022 sobre los derechos de los interesados - Derecho de acceso*. Acceso el 10 de octubre de 2024: https://www.edpb.europa.eu/system/files/2024-04/edpb_guidelines_202201_data_subject_rights_access_v2_es.pdf

Comisión Europea. 2017. *Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones relativa a la revisión intermedia de la aplicación de la Estrategia para el Mercado Único Digital. Un mercado único digital conectado para todos.* Acceso el 10 de octubre de 2024: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52017DC0228

Comisión Europea. 2018. *Comunicación de la Comisión: Inteligencia artificial para Europa*. Acceso el 10 de octubre de 2024 https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52018DC0237

Cotino, Lorenzo. 2024. «La primera sentencia del Tribunal de Justicia de la Unión Europea sobre decisiones automatizadas y sus implicaciones para la protección de datos y el Reglamento de inteligencia artificial». *Diario LA LEY* 80, *Sección Ciberderecho*. Acceso el 10 de octubre de 2024: https://diariolaley.laleynext.es/dll/2024/01/17/la-primera-sentencia-del-tribunal-de-justicia-de-la-union-europea-sobre-decisiones-automatizadas-y-sus-implicaciones-para-la-proteccion-de-datos-y-el-reglamento-de-inteligencia-artificial

DOUE. 2024. *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)*, n. 1689, de 12 de julio. Acceso el 10 de octubre de 2024: https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079

Future of Privacy Forum. 2022. *Automated decision-making under the GDPR: Practical cases from courts and data protection authorities*. Acceso el 10 de octubre de 2024: https://fpf.org/wp-content/uploads/2022/05/FPF-ADM-Report-R2-singles.pdf

Grupo de Trabajo del Artículo 29. 2017. *Directrices sobre el derecho a la portabilidad de los datos*. Acceso el 10 de octubre de 2024: https://www.aepd.es/sites/default/files/2019-09/wp242rev01-es.pdf

Grupo de Trabajo del Artículo 29. 2018. *Directrices WP 251, sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*. Acceso el 10 de octubre de 2024: https://www.aepd.es/documento/wp251rev01-es.pdf

Herrán, Isabel. 2002. *El derecho a la intimidad en la nueva ley orgánica de protección de datos personales.* Madrid: Dickinson.

Information Commissioner's Office. s.f. *How do we ensure individual rights in our AI systems?* Acceso el 10 de octubre de 2024: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-individual-rights-in-our-ai-systems/

Ministerio de Trabajo y Economía Social. 2022. *Información algorítmica en el ámbito laboral. Guía Práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral*. Acceso el 10 de octubre de 2024: https://www.mites.gob.es/ficheros/ministerio/inicio_destacados/Guia_Algoritmos_ES.pdf

Murillo de la Cueva, Pablo Lucas. 2007. «Perspectivas del derecho a la autodeterminación informativa». *IDP: Revista de Internet, derecho y política* 5: 18-32. Acceso el 10 de octubre de 2024: https://raco.cat/index.php/IDP/issue/view/6978

Parlamento Europeo. 2024. La Eurocámara aprueba una ley histórica para regular la inteligencia artificial. Notas de prensa. 13 de marzo. Acceso el 10 de octubre de 2024: https://www.europarl.europa.eu/news/es/press-room/20240308IPR19015/la-eurocamara-aprueba-una-ley-historica-para-regular-la-inteligencia-artificial

Personal Data Protection Commission Singapore. 2022. *Guía básica de anonimización*. Acceso el 10 de octubre de 2024: https://www.aepd.es/documento/guia-basica-anonimizacion.pdf

Privacy Internacional. 2017. *Data is power: Towards additional guidance on profiling and automated decision-making in the GDPR.* Acceso el 10 de octubre de 2024: https://privacyinternational.org/sites/default/files/2018-04/Data%20Is%20Power-Profiling%20and%20Automated%20Decision-Making%20in%20GDPR.pdf

Desafíos ético-jurídicos en el uso de Inteligencia Artificial para el tratamiento masivo de datos biométricos

The ethical-legal challenges in the use of Artificial Intelligence for the massive processing of biometric data

Nuria Cuadrado Gamarra
Universidad Complutense de Madrid. España
nuriacua@ucm.es
ORCiD: https://orcid.org/0000-0001-5186-988X

https://doi.org/10.18543/djhr.3199

Fecha de recepción: 31.05.2024
Fecha de aceptación: 04.08.2024
Fecha de publicación en línea: diciembre de 2024

Cómo citar / Citation: Cuadrado, Nuria. 2024. «Desafíos ético-jurídicos en el uso de la Inteligencia Artificial para el tratamiento masivo de datos biométricos». *Deusto Journal of Human Rights*, n. 14: 341-374. https://doi.org/10.18543/djhr.3199

Sumario: Introducción. 1. Conceptos previos. 2. Uso de la Inteligencia Artificial en el tratamiento de datos biométricos. 3. Recopilación y almacenamiento de datos biométricos a gran escala. 4. Desafíos éticos en la aplicación de la Inteligencia Artificial. 5. Consideraciones jurídicas y marco regulatorio. 6. Preocupaciones sobre sesgos y su impacto en los derechos fundamentales. 7. Especial referencia a los sistemas biométricos de categorización, reconocimiento de emociones y evaluación de la personalidad. 8. Implicaciones de privacidad en el escaneo de iris: análisis del caso Worldcoin. Conclusiones y recomendaciones. Bibliografía.

Resumen: El artículo examina la intersección entre la Inteligencia Artificial (IA) y el tratamiento masivo de datos biométricos, resaltando desafíos éticos y jurídicos emergentes. Comienza contextualizando el uso creciente de la IA en sistemas que emplean datos biométricos, como reconocimiento facial, huellas dactilares, escaneo del iris y voz. Se analiza la complejidad en la recopilación, almacenamiento y uso de datos biométricos a gran escala, enfocándose en la privacidad, la seguridad y el consentimiento informado. También aborda los sesgos y sus implicaciones, destacando posibles impactos discriminatorios y los desafíos para salvaguardar los derechos fundamentales en estos sistemas.

Palabras clave: Inteligencia Artificial, datos biométricos, desafíos éticos, privacidad, seguridad, consentimiento, procesamiento masivo.

Abstract: The article examines the intersection between Artificial Intelligence (AI) and the massive processing of biometric data, emphasizing emerging ethical and legal challenges. It begins by contextualizing the increasing use of AI in systems utilizing biometric data such as facial recognition, fingerprints, iris scan and voice. The complexity of collecting, storing, and utilizing biometric data on a large scale is analyzed, focusing on privacy, security, and informed consent. Additionally, it addresses biases and their implications, highlighting potential discriminatory impacts and the challenges in safeguarding fundamental rights within these systems.

Keywords: Artificial Intelligence, biometric data, ethical challenges, privacy, security, consent, massive processing.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

Introducción[1]

En la última década, se detecta un uso en prácticamente todos los sectores de la cotidiana vida individual y colectiva de una tecnología, la Inteligencia Artificial (en adelante, IA)[2], dotada de una potencia excepcional, que ha permitido contribuir a una expansión enorme y un rapidísimo desenvolvimiento de la misma. Hasta el punto de que ha terminado por convertirse en un acontecimiento social y económico, un "boom" mediático y real, una gigantesca tela de araña e instrumento que probablemente terminará siendo accesible, cómodo e imprescindible. Ingenio tecnológico que proyecta una influencia exponencial y que se encuentra en fase de crecimiento, y que, a su vez, terminará por provocar modificaciones de algunas de las nociones más básicas de la existencia.

Uno de los campos en los que la IA ha tenido un impacto significativo es en el tratamiento de datos biométricos. La capacidad de los sistemas de IA para analizar, interpretar y utilizar datos biométricos tales como reconocimiento facial, huellas dactilares, voz, iris, retina, movimientos corporales y otros atributos únicos ha revolucionado tanto la tecnología como los servicios a los que accedemos (Cotino 2022, 68; Bestard 2021, 3). De esta manera, se genera de manera permanente un imparable flujo de ingentes caudales agregados de datos biométricos, de muy diversa procedencia, naturaleza, condición y valor, confirmándose, al mismo tiempo que se multiplica, la explosión de información por sobreabundancia.

El uso generalizado de sistemas biométricos impulsados por la IA ha brindado beneficios notables en áreas como la seguridad (Barona 2024, 303), la protección de la salud, la identificación personal y la accesibilidad a diferentes servicios. Sin embargo, este progreso

[1] Este trabajo se ha realizado en el marco del Proyecto de I+D+i PID2022-136439OB-I00/ MCIN/AEI/10.13039/501100011033, Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas, financiado por el Ministerio de Ciencia y Innovación, Cofinanciado por el Fondo Europeo de Desarrollo Regional "Una manera de hacer Europa".

[2] Como es de sobra conocido, la definición del sintagma Inteligencia Artificial no resulta pacífica. Nos serviremos en este texto del artículo 3.1. del Reglamento Europeo de Inteligencia Artificial, en virtud del cual se considera IA: "un sistema basado en una máquina diseñado para funcionar con distintos niveles de autonomías, 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 información de salida, como predicciones, contenidos, recomendaciones o decisiones, que puede influir en entorno físicos o virtuales".

tecnológico conlleva una serie de desafíos éticos y jurídicos que requieren una atención minuciosa y una regulación adecuada (Aliaga y Gutiérrez 2020, 3; Etxeberria et al. 2023, 107-126).

El propósito de este artículo es analizar críticamente la intersección entre la IA y el tratamiento masivo de datos biométricos, examinando los desafíos éticos y legales emergentes en este campo. Se comenzará con una descripción detallada del uso actual de la IA en sistemas biométricos, destacando su evolución y aplicaciones. Posteriormente, se explorarán las complejidades en la recopilación, almacenamiento y uso a gran escala de datos biométricos, enfocándose en las preocupaciones éticas relacionadas con la privacidad, la seguridad o el consentimiento informado. Asimismo, se analizará la relevancia jurídica y regulatoria, evaluando la adecuación de los marcos legales existentes para abordar los desafíos planteados por la combinación de estas tecnologías avanzadas. Además, se discutirá la implementación y efectividad de las políticas actuales en diversos contextos jurídicos y culturales (Álvarez y Sanz 2021)

El análisis culminará con el estudio de un caso específico. En concreto, el controvertido uso del escaneo del iris como método de identificación biométrica que implementó Worldcoin, empresa de criptomonedas dirigida por el CEO de OpenAI Sam Altman, ofreciendo de esta manera una perspectiva práctica de los temas tratados (Andúgar 2023, Barona 2024). Finalmente, se presentarán conclusiones y recomendaciones para futuras investigaciones y políticas públicas en este ámbito.

Las aplicaciones de las tecnologías englobadas dentro del concepto amplio de IA son versátiles, abarcando diversos sectores y contextos. Estas tecnologías se utilizan desde la automatización de tareas mecánicas o rutinarias, como la gestión de comunicaciones, seguimiento de envíos, facturación y remesas, hasta operaciones más complejas, como la toma de decisiones algorítmicas en mercados financieros, clasificación de datos para la selección de personal, filtrado de contenidos, evaluación de solvencia, resolución de disputas y la interacción conversacional con humanos mediante asistentes personales, robots asistenciales y chatbots. Los riesgos asociados son variados, y las cuestiones legales que surgen difieren significativamente según el caso. Las tecnologías de reconocimiento facial y biométricas están revolucionando la forma en que se gestiona la seguridad y la privacidad en diversos contextos, desde el acceso a eventos deportivos hasta el control de la presencia en entornos laborales (Díaz Lima 2023; Espuga 2023)

Resulta necesario, al tiempo que evidente, destacar que los beneficios de la IA son extensos, tanto para el ámbito privado como público. Así, la

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

IA mejora la eficiencia de las organizaciones y sus procesos internos, dando lugar a nuevos modelos de negocio, productos y servicios. Asimismo, tiene impactos positivos para los Estados, contribuyendo a la predecibilidad en decisiones administrativas (Ponce 2024) y judiciales (Simón 2021), sistemas objetivos de selección de empleados públicos, fortalecimiento de la protección policial, identificación de delincuentes, anticipación de fraudes fiscales (Serrano 2022) o incumplimientos con entidades como la Hacienda Pública o la Seguridad Social. Sin embargo, también se han identificado efectos negativos. Las ventajas mencionadas pueden afectar derechos individuales como la intimidad, la protección de datos, la igualdad o la interdicción de la discriminación. Además, existe una importante preocupación por la posible destrucción de empleo, confusiones entre personas inocentes y delincuentes debido al mal uso de la biométrica, y un aumento de la discriminación basada en diversas circunstancias personales.

Ante estos desafíos, el 21 de abril de 2021, la Comisión Europea presentó la Propuesta de Reglamento del Parlamento Europeo y del Consejo sobre Inteligencia Artificial (Reglamento de Inteligencia Artificial o Artificial Intelligence Act), actualmente aprobado[3]. Reglamento a través del que se busca, entre otras cuestiones, establecer normas armonizadas en el ámbito de la IA para abordar los desafíos éticos y legales asociados con su uso en el contexto actual y con proyección de futuro.

1. **Conceptos previos**

Antes de adentrarnos en los detalles del almacenamiento y tratamiento de datos biométricos con sistemas de IA, entendemos que resulta esencial definir algunos conceptos clave que serán recurrentes a lo largo de este análisis. Los datos biométricos son sometidos a un "tratamiento técnico específico" según el artículo 4.14 del Reglamento General Europeo de Protección de Datos[4] (en adelante RGPD), al que en ocasiones se refiere como como tratamiento "automático" y en

[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) n. 300/2008, (UE) n. 167/2013, (UE) n. 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). Véase Simón y Cotino (2024).

[4] 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

otras como "automatizado". Este estudio excluye el simple tratamiento de vídeos y fotografías sin sistemas automatizados para extraer datos biométricos, sin cotejo con una base de datos adicional, como se indica en el Dictamen 3/2012 del G29[5]. Aunque la videovigilancia, ya sea pública o privada, plantea importantes problemas, el enfoque principal aquí se centra en el uso de componentes de IA o aprendizaje automático, que se ha vuelto cada vez más común y representa un avance significativo (Arroyo 2022).

Los sistemas de identificación biométrica automatizados, especialmente aquellos basados en IA o reconocimiento facial, tienden a facilitar la vigilancia masiva o exhaustiva. Resultan difíciles de restringir, manejan datos sensibles, las inferencias y conclusiones derivadas de ellos pueden tener un impacto considerable y la duración del tratamiento puede ser extensa. Además, existe la posibilidad de que se utilicen para fines desconocidos. Es importante señalar que la regulación actual para tratamientos "simples" de videovigilancia no es adecuada legalmente para los tratamientos con sistemas biométricos que incorporan IA y reconocimiento facial, según el Informe 31/2019 de la Agencia Española de Protección de Datos (en adelante, AEPD)[6].

En lo que concierne al uso de los datos biométricos, hay un interés particular en las técnicas de reconocimiento facial basadas en sistemas de IA. El reconocimiento facial es una funcionalidad de software que puede conectarse con diversos sistemas y combinarse con otras funciones. Aunque el reconocimiento facial automático, que implica el tratamiento de datos personales protegidos, ha generado preocupaciones, también se destaca que las técnicas de reconocimiento facial basadas en IA son emblemáticas de esta tecnología y sus riesgos. A pesar del enfoque especial en el reconocimiento facial, el régimen jurídico aplicable es el más general de los sistemas biométricos, no limitándose exclusivamente a patrones faciales.

2. **Uso de la IA en el tratamiento de datos biométricos**

La intersección en constante expansión entre la IA y el tratamiento de datos biométricos ha dejado una marca significativa en el panorama

de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos).

[5] El llamado Working Party, también conocido como G29 o Grupo de Trabajo del artículo 29, hoy Comité Europeo de Protección de Datos o EDPB por sus siglas en inglés

[6] Véase: https://www.aepd.es/documento/2019-0031.pdf

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

tecnológico contemporáneo. Este apartado se centra en una exploración exhaustiva de la aplicación de la IA en el procesamiento de datos biométricos, desglosando diversos aspectos clave que definen este paradigma tecnológico. El avance de la IA ha permitido que los sistemas de reconocimiento facial y otros métodos biométricos sean más precisos y eficientes, lo que ha llevado a su adopción en una variedad de sectores, desde la seguridad pública hasta el comercio minorista (Baz 2021).

2.1. *Evolución de la aplicación de la IA en datos biométricos*

En lo que sin duda constituye un desvergonzado análisis retrospectivo, procedemos a examinar algunos de los más destacados hitos que han propiciado mejoras sustanciales en la identificación y autenticación biométrica. Este enfoque histórico contextualiza la progresión tecnológica, resaltando cómo la IA ha impulsado avances en la precisión y eficiencia de los sistemas biométricos. Así, a lo largo de la evolución de la aplicación de la IA en el tratamiento de datos biométricos hemos sido testigos de un progreso continuo que ha transformado radicalmente la capacidad de identificación y autenticación. Desde sus primeras implementaciones hasta las más actuales, la IA ha desempeñado un papel esencial en el refinamiento de los sistemas biométricos.

En sus primeras etapas, la recopilación de datos biométricos con sistemas de IA se centraba, principalmente, en algoritmos de comparación simples para la identificación de características biométricas básicas, como huellas dactilares. A medida que la capacidad computacional aumentó, surgieron algoritmos más complejos basados en el aprendizaje automático, permitiendo una adaptación dinámica a patrones biométricos más sofisticados. Con el tiempo, los algoritmos de reconocimiento facial impulsados por la IA han experimentado avances notables. La capacidad de la IA para analizar de manera eficiente las características faciales, considerando aspectos como la variabilidad en expresiones y ángulos de captura, ha mejorado significativamente. Además, el aprendizaje profundo ha permitido modelos más precisos y robustos, contribuyendo a la identificación facial en condiciones diversas. En paralelo, el procesamiento de lenguaje natural ha influido en la interpretación y autenticación de características biométricas como la voz y los patrones de escritura. Los algoritmos basados en esta rama de la IA han logrado una comprensión más profunda de las variaciones en el habla y en la

escritura, mejorando la precisión y la adaptabilidad de los sistemas biométricos relacionados.

La evolución de la aplicación de la IA en datos biométricos no solo se ha traducido en avances tecnológicos, sino que también ha impactado en la accesibilidad y la eficiencia de los sistemas. La capacidad de aprendizaje continuo de la IA ha permitido adaptarse a nuevos patrones biométricos sin intervención humana constante, facilitando la implementación en entornos dinámicos. En conjunto, la evolución en la aplicación de la IA en datos biométricos refleja una trayectoria ascendente, desde métodos iniciales hasta la adopción de técnicas más avanzadas, marcando así una era donde la IA y los datos biométricos convergen para ofrecer soluciones cada vez más precisas y adaptativas en el ámbito de la identificación y autenticación.

2.2. *Aprendizaje automático y procesamiento de lenguaje natural en datos biométricos*

La fusión de la IA con el tratamiento de datos biométricos se ha potenciado significativamente a través del aprendizaje automático y el procesamiento de lenguaje natural (en adelante, PLN). El aprendizaje automático ha revolucionado la interpretación de datos biométricos al permitir que los sistemas se adapten y evolucionen con la experiencia. En el contexto biométrico, algoritmos de aprendizaje automático, como redes neuronales y máquinas de soporte vectorial, han demostrado ser especialmente eficaces en el reconocimiento facial y de voz. Estos algoritmos pueden aprender patrones complejos y variaciones, mejorando la precisión y la capacidad de adaptación de los sistemas biométricos. De la misma manera, el PLN ha ampliado la aplicación de datos biométricos más allá de las tradicionales huellas dactilares y reconocimiento facial. En el ámbito biométrico, el PLN se aplica a la identificación de patrones en la voz y en los patrones de escritura. Los algoritmos de PLN pueden analizar la entonación, el ritmo y otros aspectos lingüísticos, mejorando la precisión en la autenticación biométrica basada en la voz. Además, el PLN permite la interpretación de patrones en la escritura, contribuyendo así a una autenticación más holística.

La integración de aprendizaje automático y PLN en sistemas biométricos ha impulsado la capacidad de interpretar datos complejos y variados. En el reconocimiento facial, por ejemplo, los algoritmos pueden aprender a reconocer expresiones faciales específicas y adaptarse a diferentes condiciones de iluminación. En el

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

reconocimiento de voz, la combinación de PLN con aprendizaje automático permite una interpretación más precisa y contextual de los patrones vocales, mejorando la autenticación biométrica.

2.3. *Aplicaciones prácticas en seguridad, salud y otros sectores*

Si analizamos las aplicaciones prácticas que han surgido de la convergencia entre la IA y los datos biométricos, debemos detenernos en examinar casos de implementaciones exitosas en sectores clave, como la seguridad, la salud y el ámbito empresarial, identificando beneficios tangibles y potenciales riesgos asociados con la adopción masiva de estas tecnologías. La confluencia de la IA y el tratamiento de datos biométricos ha desencadenado una proliferación de aplicaciones prácticas, marcando un impacto significativo en diversos sectores. A continuación, trataremos de, brevemente, destacar algunos casos específicos de implementaciones exitosas en áreas clave, resaltando tanto los beneficios tangibles como los riesgos asociados.

Uno de los sectores más prominentes que ha experimentado la influencia de la IA en datos biométricos es el de la seguridad. La aplicación de reconocimiento facial y huellas dactilares ha mejorado la autenticación en sistemas de acceso, desde desbloqueo de dispositivos hasta control de acceso a instalaciones críticas. La IA permite una identificación más precisa y eficiente, fortaleciendo la seguridad en entornos sensibles.

En el ámbito de la salud, la IA ha encontrado aplicación en la identificación biométrica para garantizar la integridad del historial médico y la autenticación de pacientes. Sistemas biométricos basados en el reconocimiento de iris, huellas dactilares o incluso patrones de voz han mejorado la precisión y seguridad en el acceso a información médica sensible, contribuyendo a una atención médica más personalizada y segura.

La implementación de la IA en la gestión de identidad empresarial ha facilitado el control de accesos y la autenticación de empleados. El reconocimiento facial y otras formas de autenticación biométrica han simplificado los procesos de registro y asegurado la integridad de los sistemas de seguridad corporativos. Esto no solo optimiza la eficiencia, sino que también mitiga riesgos asociados con el acceso no autorizado.

Aunque las aplicaciones prácticas de la IA en datos biométricos ofrecen beneficios notables, también plantean desafíos y riesgos. La vulnerabilidad a ataques de manipulación de datos, la posibilidad de discriminación inherente a ciertos algoritmos, y la preocupación por la

privacidad son aspectos críticos que deben abordarse para garantizar un despliegue ético y seguro de estas tecnologías.

2.4. *Desafíos éticos y de privacidad en el uso y diseño de la IA con datos biométricos*

El análisis crítico de los desafíos éticos y de privacidad que emergen con la integración más profunda de la IA en el ámbito biométrico aborda casos específicos de controversias éticas, centrándose particularmente en el reconocimiento facial y otros métodos biométricos. Este análisis proporciona una visión matizada de las preocupaciones éticas y de privacidad que requieren una atención cuidadosa en el desarrollo y aplicación de sistemas biométricos impulsados por la IA. La convergencia entre la IA y los datos biométricos, si bien ha brindado avances significativos, plantea desafíos éticos y de privacidad que requieren una consideración cuidadosa y un enfoque equilibrado.

Uno de los desafíos éticos más prominentes radica en la presencia de sesgos y discriminación en los sistemas biométricos impulsados por la IA. Estos sistemas pueden exhibir sesgos inherentes a los conjuntos de datos utilizados para su entrenamiento, lo que puede resultar en identificaciones incorrectas y discriminación, especialmente hacia grupos étnicos minoritarios. La necesidad de abordar estos sesgos para garantizar la equidad y evitar consecuencias discriminatorias se presenta como un imperativo ético (Cotino 2023a; Díaz Lima 2023).

La recopilación masiva de datos biométricos plantea interrogantes éticos en relación con el consentimiento informado y la autonomía del individuo. La falta de comprensión completa sobre cómo se utilizarán y compartirán estos datos, así como la posibilidad de usos no autorizados, plantea preocupaciones sobre la transparencia y el control del individuo sobre su información biométrica. De este modo, garantizar un proceso de consentimiento claro y transparente se erige como un principio ético fundamental.

La seguridad de los datos biométricos frente a amenazas y ataques cibernéticos constituye otro desafío ético crucial. La posibilidad de manipulación o robo de datos biométricos plantea riesgos significativos para la privacidad y la seguridad de los individuos. Se requieren medidas rigurosas para salvaguardar la integridad de estos datos y mitigar las amenazas potenciales.

Asimismo, la ausencia de un marco regulatorio robusto y la asignación clara de responsabilidades éticas en el uso de la IA con

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

datos biométricos son desafíos adicionales. La falta de normativas claras puede dar lugar a prácticas no éticas o incluso ilegales en la recopilación y tratamiento de datos biométricos. Establecer un marco regulatorio sólido y fomentar la responsabilidad ética en todas las fases de desarrollo y aplicación de estas tecnologías se presenta como una necesidad imperante.

El uso de datos biométricos en el ámbito jurídico y penal plantea importantes retos y consideraciones éticas, como se discute en diversas investigaciones recientes (Etxeberria Guridi et al. 2023; Flórez y Camelo 2023). Estas tecnologías no solo afectan la privacidad de los individuos, sino que también tienen implicaciones legales significativas (Garriga et al. 2023; Garrós 2021). La discriminación algorítmica y su impacto en la dignidad de la persona y los derechos humanos es otro aspecto crucial (Iturmendi 2023).

3. **Recopilación y almacenamiento de datos biométricos a gran escala**

La recopilación masiva de datos biométricos plantea desafíos significativos en términos de privacidad y seguridad. Las organizaciones deben asegurar que los datos se almacenen de manera segura y se utilicen de manera ética (Bestard 2021) La recopilación de datos biométricos a gran escala involucra la captura y análisis de un volumen masivo de información única para cada individuo. Esto incluye, entre otros, el reconocimiento facial, huellas dactilares y patrones de voz. La complejidad radica en la variedad de fuentes y métodos de recopilación, desde sistemas de videovigilancia hasta dispositivos móviles, lo que exige una atención meticulosa a la precisión y la integridad de los datos.

La escala masiva de la recopilación de datos biométricos plantea cuestiones éticas fundamentales. La obtención de datos sin el conocimiento o el consentimiento informado de los individuos puede comprometer la privacidad o la autonomía. La transparencia en los procesos de recopilación y el respeto por los derechos individuales se tornan imperativos éticos en este contexto. El uso de IA en el reconocimiento facial y otros sistemas biométricos ha generado preocupaciones sobre la invasión de la privacidad, la vigilancia masiva y la posibilidad de errores y sesgos en los algoritmos (Castellanos 2023)

El almacenamiento seguro de datos biométricos es esencial para mitigar riesgos asociados con la seguridad y la privacidad. La vulnerabilidad de estos datos ante amenazas cibernéticas exige

medidas robustas de seguridad, como cifrado avanzado y protocolos de acceso controlado. Además, se debe garantizar la anonimización efectiva para proteger la identidad de los individuos, minimizando así el riesgo de uso indebido. Asimismo, la retención y eliminación de datos biométricos plantea desafíos éticos sobre la duración justificada y necesaria de la retención. Establecer políticas claras que rijan la retención y el período de conservación es esencial para evitar la acumulación innecesaria de información y garantizar la gestión ética de los datos biométricos a lo largo del tiempo (Flórez y Camelo 2023; Sempere 2020).

La ausencia de un marco regulatorio sólido y la falta de principios éticos claros en la recopilación y almacenamiento de datos biométricos pueden generar prácticas inconsistentes y riesgos para los individuos. Establecer directrices éticas y reglamentaciones que guíen la gestión de datos biométricos a gran escala es esencial para garantizar una práctica ética y legal en este ámbito. A este respecto, debemos destacar el cambio significativo que introduce el nuevo Reglamento de IA (RIA) y que posteriormente desarrollaremos. Asimismo, se analizará la relevancia jurídica y regulatoria, evaluando la adecuación de los marcos legales existentes para abordar los desafíos planteados por la combinación de IA y datos biométricos (Razquin 2022; Romano 2023).

En el mundo democrático, el uso de sistemas biométricos de identificación ha generado numerosos conflictos jurisdiccionales, destacando experiencias en el Reino Unido, Alemania, Países Bajos, Italia, Suecia, Buenos Aires, Brasil y España. Se han implementado proyectos polémicos, como el *AFR Locate* en el Reino Unido[7] y sistemas biométricos en eventos en Países Bajos. Asimismo, han surgido controversias legales y éticas en Alemania, donde el Bundesverfassunggericht declaró la inconstitucionalidad de dos normas que permitían el uso de sistemas automatizados en el ámbito policial (Cotino 2023b).

En Buenos Aires un sistema de cámaras establecido con el objetivo de garantizar la seguridad, identificó erróneamente a personas por lo que un tribunal suspendió su uso[8], mientras que, en Brasil, un juzgado ordenó la suspensión del uso de un sistema de control biométrico

[7] El once de agosto de 2020 una *High Court* del Reino Unido declaró la ilegalidad, por ausencia de transparencia y eficacia, de un sistema de reconocimiento facial empleado por la Policía de Gales del Sur. Véase: extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.judiciary.uk/wpcontent/uploads/2020/08/R-Bridges-v-CC-South-Wales-ors-Judgment.pdf.

[8] Véase: https://www.accessnow.org/buenos-aires-y-sao-paulo-suspenden-reconocimiento-facial.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

implementado por la empresa ViaQuatro, concesionaria del metro de Sao Paulo[9]. En España, el uso de sistemas biométricos en el ámbito privado ha generado conflictos legales, con sanciones a empresas como Mercadona[10]. Además, se menciona la aplicación de sistemas biométricos en el ámbito educativo y privado. Por último, sirva como ejemplo, y probablemente como advertencia de los desafíos a los que nos enfrentamos, el uso de sistemas de reconocimiento facial establecido recientemente por el Gobierno iraní en orden a controlar el uso del velo incluso cuando las mujeres se encuentran dentro de su vehículo (Harari 2024, 296) o los sistemas de reconocimiento facial que incorporan IA en China (Han 2020).

Se destaca que, más allá de la identificación, los sistemas biométricos inteligentes permiten el reconocimiento de emociones y la evaluación de personalidades. La Comisión Europea financió un proyecto, *Intelligent Portable Control System* (iBorderCtrlla)[11], con el objeto de aplicar la IA a los ámbitos de la migración y el asilo que generó encontradas reacciones en la sociedad civil y en la doctrina más autorizada, dado el alto grado de afectación a un buen número de derechos fundamentales. Grandes colosos empresariales, como Microsoft y Meta-Facebook, han retirado o modificado sus sistemas de reconocimiento facial y biométrico en respuesta a preocupaciones éticas. Aunque estos casos son significativos en el mundo democrático, se plantea la incertidumbre sobre el uso de sistemas biométricos en contextos más vinculados a la defensa y seguridad nacional, especialmente en China.

4. **Desafíos éticos en la aplicación de la IA**

La presencia de sesgos en los algoritmos de IA, especialmente aquellos entrenados en conjuntos de datos sesgados, puede generar discriminación. Esto se traduce en decisiones y recomendaciones que pueden tener consecuencias perjudiciales para ciertos grupos étnicos, de género o socioeconómicos. Abordar estos sesgos y garantizar la equidad en la aplicación de la IA son imperativos éticos.

Por otro lado, la opacidad en el funcionamiento interno de algunos modelos de IA plantea desafíos éticos relacionados con la falta de

[9] Véase: https://www.accessnow.org/press-release/sao-paulo-tribunal-prohibe-camaras-de-reconocimiento-facial-en-el-metro/.

[10] Procedimiento PS/00120/2021 de la AEPD, de 27 de julio de 2021.

[11] Véase: https://www.iborderctrl.eu/.

transparencia y explicabilidad. La dificultad para comprender cómo toma decisiones un sistema de IA puede socavar la confianza de los usuarios y dificultar la rendición de cuentas. Establecer prácticas transparentes y asegurar la explicabilidad de los modelos son aspectos clave para abordar este desafío.

Asimismo, la recopilación masiva de datos para entrenar modelos de IA presenta preocupaciones éticas significativas en términos de privacidad y seguridad. La posibilidad de identificar patrones sensibles en datos personales y la amenaza de brechas de seguridad plantean riesgos para la privacidad individual. Garantizar prácticas de manejo de datos éticas y protocolos de seguridad robustos es esencial para proteger la información personal.

La automatización impulsada por la IA tiene el potencial de alterar significativamente el panorama laboral, lo que genera desafíos éticos relacionados con el desempleo y los cambios socioeconómicos. La necesidad de reentrenamiento y adaptación de las habilidades laborales se convierte en un imperativo ético para mitigar los impactos negativos en los trabajadores afectados.

Definir la responsabilidad y establecer mecanismos efectivos de rendición de cuentas en el desarrollo y aplicación de sistemas de IA es un desafío ético clave. En situaciones donde los sistemas de IA toman decisiones críticas, la asignación clara de responsabilidades se vuelve esencial para abordar posibles consecuencias adversas y garantizar una toma de decisiones ética.

Por último, el uso de IA en contextos militares plantea preocupaciones éticas sobre la potencial falta de control humano, la escalada de conflictos y el impacto desigual en regiones y poblaciones. Abordar estos desafíos implica establecer límites éticos claros en el desarrollo y aplicación de tecnologías de IA en el ámbito militar.

5. **Consideraciones jurídicas y marco regulatorio**

La legislación en torno a la IA y los datos biométricos está en constante evolución. Es crucial que las leyes sean actualizadas para proteger adecuadamente los derechos de los individuos sin sofocar la innovación tecnológica (Castellanos 2023). La recopilación y procesamiento de grandes cantidades de datos en la aplicación de la IA requieren un estricto cumplimiento de las leyes de protección de datos y privacidad. El marco regulatorio debe establecer principios claros sobre cómo se pueden recopilar, almacenar y utilizar los datos para garantizar que se respeten los derechos fundamentales de privacidad de los individuos.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

Las leyes y regulaciones deben abordar la necesidad de transparencia y explicabilidad en los sistemas de IA. Esto implica la obligación de que las organizaciones proporcionen información clara sobre cómo funcionan sus algoritmos, especialmente en casos donde las decisiones afectan a los individuos. La capacidad de entender y cuestionar las decisiones de la IA es fundamental para la rendición de cuentas.

Establecer la responsabilidad legal en casos de decisiones erróneas o consecuencias adversas causadas por sistemas de IA es asimismo esencial. El marco regulatorio debe definir claramente quién es responsable en diferentes fases del ciclo de vida de la IA, ya sea en el diseño, entrenamiento o implementación. Además, se deben establecer mecanismos para la rendición de cuentas en situaciones donde los sistemas de IA toman decisiones cruciales.

Las leyes y regulaciones deben abordar los aspectos de seguridad cibernética en el contexto de la IA. Esto implica establecer estándares de seguridad para proteger los sistemas de IA contra amenazas y ataques cibernéticos. La integridad y la confidencialidad de los datos procesados por sistemas de IA deben ser una prioridad legal. Las normas deben también abordar la discriminación y los sesgos en los algoritmos de IA. Esto implica la implementación de medidas para garantizar que los sistemas de IA no perpetúen o amplifiquen sesgos existentes en la sociedad. La prohibición de discriminación injusta basada en raza, género u otras características protegidas debe ser una consideración central.

El marco legal debe abordar cuestiones relacionadas con la propiedad intelectual y los derechos de autor en el desarrollo de la IA. Esto incluye la definición de la propiedad de los modelos de IA, algoritmos y resultados generados por sistemas de IA, así como la protección de los derechos de los desarrolladores y creadores involucrados en proyectos de IA. Por otro lado, las leyes y regulaciones deben incluir disposiciones éticas para guiar la investigación y el desarrollo de la IA. Esto implica establecer límites éticos claros en la experimentación y el uso de datos, así como la consideración de posibles impactos éticos en la sociedad. Este análisis destaca la importancia de un marco regulatorio integral que aborde las complejidades jurídicas asociadas con la aplicación de la IA, asegurando así un desarrollo ético, responsable y conforme a los principios legales fundamentales.

La literatura existente muestra una diversidad de enfoques y regulaciones en distintos países, lo cual se refleja en el análisis comparativo de Flórez y Camelo (2023) sobre las tecnologías de

reconocimiento facial en Colombia. Además, los aspectos técnicos y éticos del tratamiento de datos biométricos se han debatido ampliamente en Europa (Garrós 2021). También se ha destacado la necesidad de abordar la discriminación algorítmica en estos contextos (Iturmendi 2023)

A nivel europeo y en el momento actual, son escasas las referencias explícitas que encontramos respecto al tratamiento masivo de datos biométricos, destacando el Reglamento (CE) n. 444/2009 del Parlamento Europeo y del Consejo, de 6 de mayo de 2009, por el que se modifica el Reglamento (CE) n. 2252/2004 del Consejo sobre normas para las medidas de seguridad y datos biométricos en los pasaportes y documentos de viaje expedidos por los Estados miembros; el 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 (RGPD); la Decisión de Ejecución (UE) 2023/1795 de la Comisión de 10 de julio de 2023 relativa a la adecuación del nivel de protección de los datos personales en el Marco de Privacidad de Datos UE-EE.UU. con arreglo al Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo [notificada con el número C(2023) 4745] y el Reglamento (UE) 2022/991 del Parlamento Europeo y del Consejo, de 8 de junio de 2022, por el que se modifica el Reglamento (UE) 2016/794 en lo que se refiere a la cooperación de Europol con entidades privadas, el tratamiento de datos personales por Europol en apoyo de investigaciones penales y el papel de Europol en materia de investigación e innovación. En cuanto a la regulación nacional, debemos destacar la Resolución de 6 de abril de 2016, del Servicio Público de Empleo Estatal, por la que se aprueba el sistema de firma electrónica mediante captura de firma digitalizada con datos biométricos para relacionarse presencialmente con el Servicio Público de Empleo Estatal; la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales; la Orden ETD/465/2021, de 6 de mayo, por la que se regulan los métodos de identificación remota por vídeo para la expedición de certificados electrónicos cualificados, y la Ley Orgánica 7/2021, de 26 de mayo, de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

6. **Preocupaciones sobre sesgos y su impacto en los derechos fundamentales**

La presencia de sesgos en sistemas de IA plantea inquietudes significativas, especialmente en relación con la posible afectación de los derechos fundamentales de los individuos. El uso diversificado de tecnologías biométricas, especialmente el reconocimiento facial, tiene el potencial de afectar prácticamente todos los derechos fundamentales de las personas (Sebé y Santilari 2022; Garrido 2023). La Agencia de la Unión Europea para los Derechos Fundamentales (FRA) destaca la implicación de derechos como la dignidad humana, el respeto a la vida privada, la protección de datos personales, la no discriminación, los derechos del niño y de los mayores, la libertad de reunión y asociación, la libertad de expresión, el derecho a una buena administración, y el derecho a un recurso efectivo ante la ley y a un juicio justo.

La sociedad democrática se ve amenazada por el control constante y la posible identificación en situaciones cotidianas, desde comprar pan hasta participar en manifestaciones, actividades políticas o religiosas, centros de salud, asistencia social, educativos, entre otros. El RIA reconoce que estos sistemas pueden generar una sensación de vigilancia constante, disuadiendo indirectamente el ejercicio de la libertad de reunión y otros derechos fundamentales. Además, estos sistemas ponen en entredicho la presunción de inocencia al considerar a todos como sospechosos, dificultando la defensa de aquellos que resultan positivos en estas estructuras. Los errores, sesgos y discriminaciones, aunque inadvertidos, son posibles dado que estos sistemas son probabilísticos y aplicados a grandes poblaciones. Los derechos asociados a la privacidad, intimidad y protección de datos se ven particularmente afectados, ya que la simple captación de datos, incluso si se eliminan inmediatamente después de su comparación, resulta impactante. La utilización y conservación posterior de estos datos procesados intensifican la afectación a estos derechos y aumentan el riesgo de un uso indebido.

El reconocimiento facial y estas tecnologías constituyen una amenaza directa a los derechos fundamentales, y su impacto va más allá de afectaciones individuales, alcanzando a la sociedad democrática en su conjunto. La respuesta jurídica no puede limitarse a aplicar técnicas específicas para afectaciones individuales; se requieren nuevas técnicas de cumplimiento normativo, responsabilidad reactiva, análisis multirriesgos y protección colectiva de derechos para abordar de manera integral estos desafíos éticos y legales. El sesgo en los sistemas

de IA puede llevar a decisiones injustas y discriminatorias. Es fundamental que los desarrolladores y reguladores trabajen juntos para minimizar estos riesgos (Castellanos 2024). La experiencia demuestra que la existencia de sesgos en los algoritmos de IA puede resultar en discriminación injusta hacia ciertos grupos. Esto podría manifestarse en decisiones discriminatorias en áreas como el empleo, la vivienda, o el crédito, entre otros. La discriminación basada en sesgos en sistemas de IA amenaza directamente el derecho a la igualdad y no discriminación, fundamentales en los derechos humanos.

Los sesgos en los sistemas de IA pueden afectar el derecho a la privacidad al influir en la recopilación y procesamiento de datos de manera desigual. Si los algoritmos muestran sesgos en la identificación de patrones o en la toma de decisiones, esto puede traducirse en un tratamiento diferenciado y, en última instancia, en la vulneración del derecho a la privacidad.

La aplicación de sistemas de IA en procesos legales, como la toma de decisiones judiciales o la evaluación de riesgos penales, podría verse afectada por sesgos inherentes. Esto plantea preocupaciones sobre el derecho a un juicio justo y a la igualdad ante la ley, ya que las decisiones automatizadas pueden favorecer o perjudicar a ciertos grupos de manera no equitativa.

Si los algoritmos de IA muestran sesgos, el acceso a oportunidades fundamentales como empleo, educación y atención médica puede volverse inequitativo. Esto amenaza el derecho a la igualdad de oportunidades, ya que las decisiones basadas en sesgos pueden perpetuar desigualdades existentes y limitar el acceso a recursos esenciales.

Los sesgos en los algoritmos que determinan qué contenido se muestra a los usuarios pueden tener un impacto en la libertad de expresión e información. Si los usuarios son expuestos a información sesgada o limitada, esto puede afectar su capacidad para formar opiniones informadas y participar plenamente en el discurso público.

En entornos laborales donde se utilizan sistemas de IA para la toma de decisiones relacionadas con el empleo, los sesgos pueden influir en la contratación, la promoción y la evaluación del desempeño. Esto plantea preocupaciones sobre el derecho a un ambiente laboral justo y la igualdad en el trato en el ámbito laboral.

Abordar las preocupaciones sobre sesgos en la IA se convierte en una prioridad crucial para salvaguardar los derechos fundamentales y garantizar que la aplicación de estas tecnologías sea justa, equitativa y respetuosa con los principios fundamentales de la normativa que reconoce y garantiza los derechos humanos.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

7. **Especial referencia a los sistemas biométricos de categorización, reconocimiento de emociones y evaluación de la personalidad**

En el ámbito de las tecnologías biométricas e IA, la lectura de datos faciales, indicadores sanguíneos, pulsación de teclas y otros elementos está adquiriendo una relevancia creciente. Aunque estos datos son universales y singularizan a la persona, su uso se ha expandido más allá de la mera identificación. En la propuesta del RIA de abril de 2021, se define un "sistema de reconocimiento de emociones" como un sistema de IA destinado a identificar o inferir emociones a partir de datos biométricos. Se sugiere que esta definición debería incluir los "pensamientos" a través de interfaces cerebro-ordenador. Además, se introduce el concepto de "sistema de categorización biométrica", destinado a asignar a las personas a categorías específicas, como sexo, edad, color de pelo, entre otros, basándose en datos biométricos.

Estos sistemas, según la propuesta, permiten una evaluación a gran escala que se asemeja a la capacidad de un psicólogo para interpretar emociones, detectar veracidad en manifestaciones y prever comportamientos futuros. Además, posibilitan la rápida categorización de conjuntos de personas con afinidades específicas. A lo largo de los años, se han empleado estos sistemas en el control de fronteras, como el Agente Virtual Automatizado para la Evaluación de la Verdad en Tiempo Real (AVATAR) en los EE.UU. Asimismo, estos sistemas plantean cuestiones éticas adicionales debido a su capacidad para inferir características personales profundas a partir de datos biométricos, lo que podría ser explotado de manera indebida (Díaz 2023).

A pesar de sus beneficios potenciales, se han identificado preocupaciones (Andúgar 2023; Castellanos 2023), y varias instituciones han señalado el peligro que representan estos sistemas biométricos inteligentes. Empresas como Microsoft y Meta-Facebook han tenido en cuenta estas inquietudes, retirando o modificando sus sistemas de reconocimiento facial y de emociones. Aunque estos sistemas generan inquietudes, la regulación actual, como el RGPD y el RIA, presenta limitadas precauciones y garantías, especialmente cuando los datos biométricos no se utilizan con fines de identificación. La regulación futura y las llamadas a la prohibición o regulación más estricta por parte de importantes instituciones subrayan la necesidad de abordar de manera más integral estos desafíos éticos y legales asociados con la IA y las tecnologías biométricas.

8. **Implicaciones de privacidad en el escaneo de iris: análisis del caso worldcoin**

El escaneo de iris ha emergido como una tecnología biométrica innovadora que permite la identificación precisa de individuos mediante el análisis de los patrones únicos presentes en el iris del ojo. Esta tecnología ha encontrado una amplia gama de aplicaciones, desde el control de acceso en instalaciones de alta seguridad hasta la autenticación en dispositivos móviles. Sin embargo, a medida que el escaneo de iris se integra cada vez más en nuestras vidas cotidianas, también surgen importantes interrogantes sobre la privacidad y la protección de datos personales. En este contexto, es crucial entender cómo la regulación de la UE, como el RGPD, maneja la protección de estos datos sensibles (González Calvo 2022).

Worldcoin ha sido objeto de controversia por su método de escaneo del iris, que muchos consideran una invasión significativa de la privacidad. Este caso será examinado para ilustrar los desafíos prácticos y éticos en la implementación de tecnologías biométricas (Díaz 2023). El escaneo del iris como método de identificación implica el tratamiento de datos biométricos, que son considerados datos personales y están sujetos a regulaciones estrictas en España para proteger la privacidad y los derechos de los individuos. Las leyes y regulaciones aplicables, como la Ley Orgánica 7/2021, establecen condiciones específicas para el tratamiento, uso y protección de estos datos, asegurando que cualquier uso del escaneo del iris cumpla con los principios de protección de datos personales y respete los derechos fundamentales. El uso del escaneo de iris para identificar a las personas está permitido en España bajo ciertas condiciones específicas, principalmente en contextos relacionados con la aplicación de la ley. Además, el escaneo de iris es reconocido como un método efectivo y común de identificación biométrica en la Unión Europea, lo que refuerza su relevancia y adopción en diversos países miembros.

Debemos destacar que el escaneo de iris representa un hito significativo en el campo de la tecnología biométrica debido a su capacidad para proporcionar una identificación precisa y única de individuos basada en características biológicas internas del cuerpo humano. A diferencia de otras formas de autenticación, como contraseñas o tarjetas de identificación, que pueden ser olvidadas, robadas o falsificadas, el escaneo de iris ofrece un nivel de seguridad y fiabilidad excepcionales. Una de las razones clave de la importancia del escaneo de iris en la tecnología biométrica radica en la singularidad y estabilidad del iris humano. Cada iris presenta patrones únicos que se

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

desarrollan durante la gestación y permanecen prácticamente inalterados a lo largo de la vida de un individuo. Esta característica inherente hace que el escaneo de iris sea altamente confiable para la identificación biométrica, lo que lo convierte en una opción atractiva para una variedad de aplicaciones críticas, como la seguridad nacional, el control de acceso a instalaciones sensibles y la autenticación en sistemas de información.

Otra razón importante es la rapidez y facilidad de uso del escaneo de iris. A diferencia de otros métodos biométricos que pueden requerir interacciones físicas más complejas, como la huella dactilar o el reconocimiento facial, el escaneo de iris puede realizarse de manera rápida y sin contacto directo, lo que lo convierte en una opción conveniente para una amplia gama de situaciones y entornos. Además, el escaneo de iris ofrece un alto nivel de precisión y resistencia a la falsificación. Los patrones del iris son extremadamente detallados y difíciles de reproducir, lo que hace que sea extremadamente difícil para los impostores engañar al sistema. Esto lo convierte en una herramienta valiosa para combatir la suplantación de identidad y otros tipos de fraude.

En definitiva, el escaneo de iris juega un papel fundamental en la tecnología biométrica al ofrecer una combinación única de precisión, seguridad y facilidad de uso. Su capacidad para proporcionar identificaciones confiables y seguras lo convierte en una herramienta invaluable en una amplia variedad de aplicaciones, desde la seguridad nacional hasta la gestión de identidad en entornos empresariales y gubernamentales.

8.1. *Regulación*

La Ley Orgánica 3/2018[12] y la Ley Orgánica 7/2021[13] son fundamentales para entender el tratamiento de datos biométricos, como el escaneo del iris, en España. La Ley Orgánica 3/2018 establece que la protección de las personas físicas en relación con el tratamiento de datos personales es un derecho fundamental, garantizando a los

[12] Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales. (Ley Orgánica 3/2018, de 5 de diciembre) BOE-A-2018-16673

[13] Ley Orgánica de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales. (Ley Orgánica 7/2021, de 26 de mayo) BOE-A-2021-8806

individuos el control sobre sus datos personales. Esta ley se centra en salvaguardar la privacidad y los derechos de los ciudadanos en el ámbito digital y biométrico. La Ley Orgánica 7/2021 especifica que los datos biométricos se consideran una categoría especial de datos personales cuando se utilizan para identificar de manera unívoca a una persona. Esto implica que el tratamiento de estos datos está sujeto a regulaciones estrictas para proteger la privacidad de los individuos y asegurar que se manejen de manera ética y legal. Además, el Artículo 18.4 de la Constitución Española establece límites al uso de la informática para proteger la intimidad personal, lo cual es particularmente relevante en el contexto del escaneo de iris.

La Ley Orgánica 7/2021, de 26 de mayo, sobre la protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales, establece que los datos biométricos, como el escaneo de iris, pueden ser utilizados legalmente para identificar de manera unívoca a una persona física en contextos específicos de aplicación de la ley. Esta ley, de ámbito nacional, tiene preeminencia sobre otras normativas autonómicas o más antiguas en caso de conflicto, dada su jerarquía y actualidad.

Normativa que, en conjunto, proporciona un marco robusto para el tratamiento de datos biométricos, asegurando que su uso se realice de manera responsable y en consonancia con los derechos fundamentales de los individuos. La regulación estricta y clara es esencial para equilibrar los beneficios del escaneo de iris como una herramienta de seguridad y autenticación, con la necesidad de proteger la privacidad y otros derechos fundamentales en una sociedad cada vez más digitalizada.

La Resolución n. PS-00120-2021 de la Agencia Española de Protección de Datos distingue entre identificación biométrica remota y autenticación biométrica, aclarando que la identificación biométrica remota, que puede incluir el escaneo del iris, debe realizarse bajo condiciones que respeten la normativa de protección de datos y los derechos individuales. Esta resolución establece directrices claras para el uso de tecnologías biométricas, garantizando que su implementación se alinee con las leyes de privacidad y protección de datos vigentes, asegurando así la protección de los derechos fundamentales de los individuos.

Los materiales secundarios, como los principios del Reglamento 2016/679/UE (RGPD) y las claves jurídicas sobre el registro de jornada, proporcionan contexto adicional sobre cómo se deben manejar los datos biométricos en diferentes contextos, como el laboral, y subrayan

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

la necesidad de garantías adecuadas para proteger los derechos fundamentales al utilizar tecnologías como el escaneo del iris. Específicamente, el documento titulado *Los sistemas biométricos de reconocimiento facial en la Unión Europea en el marco del desarrollo de la Inteligencia Artificial*, de fecha 1 de junio de 2023, destaca que el escaneo de iris es un método común y altamente efectivo para la identificación biométrica. Este documento subraya la eficacia casi absoluta del uso del iris como dato biométrico para acceder a lugares o dispositivos, apoyando la proposición de su uso extendido en contextos donde se requiere una identificación segura y precisa. Estos materiales complementarios refuerzan la importancia de aplicar regulaciones estrictas y garantías adecuadas para proteger la privacidad y los derechos de los individuos en el uso de tecnologías biométricas avanzadas.

La Sentencia 119/2022 del Tribunal Constitucional[14] es crucial en el marco legal español, ya que confirma que los datos biométricos, incluido el escaneo del iris, son considerados datos personales y, por lo tanto, están protegidos bajo la legislación de protección de datos personales en España. Esta sentencia no solo reafirma la clasificación de los datos biométricos como datos personales, sino que también subraya la obligación de cumplir con todas las normativas de protección de datos al tratarlos. El Tribunal Constitucional, en esta sentencia, enfatiza que el tratamiento de datos biométricos debe realizarse en estricto cumplimiento con la Ley Orgánica 3/2018 de Protección de Datos Personales y Garantía de los Derechos Digitales (LOPDGDD) y el RGPD. Esto incluye principios fundamentales como la licitud, lealtad y transparencia, la limitación de la finalidad, la minimización de datos, la exactitud, la limitación del plazo de conservación, la integridad y confidencialidad, y la responsabilidad proactiva.

La sentencia también aborda la necesidad de un consentimiento explícito e informado por parte de los individuos cuyo iris se va a escanear, así como la obligación de implementar medidas de seguridad adecuadas para proteger estos datos sensibles de accesos no

[14] Pleno Sentencia 119/2022, de 29 de septiembre de 2022. Recurso de amparo 7211-2021. Promovido por Saltoki Araba, S.A., en relación con las resoluciones dictadas por las salas de lo social del Tribunal Supremo y del Tribunal Superior de Justicia del País Vasco en proceso por despido. Vulneración de los derechos a la utilización de los medios de prueba pertinentes y a un proceso con todas las garantías, en conexión con el derecho a la tutela judicial efectiva: resoluciones judiciales que, sin un verdadero motivo jurídico, declaran improcedente la prueba videográfica aportada por la empresa y admitida en la instancia. Voto particular. BOE n. 262, de 1 de noviembre de 2022, páginas 149412 a 149442.

autorizados y usos indebidos. Esta protección se extiende a todos los ámbitos donde se pueda utilizar el escaneo de iris, incluyendo tanto el sector público como el privado. Además, la STC 119/2022 recalca la importancia de que las organizaciones que manejen datos biométricos establezcan políticas claras y transparentes sobre su tratamiento, incluyendo información detallada sobre los fines del procesamiento, los derechos de los individuos respecto a sus datos, y los procedimientos para ejercer estos derechos. Esto refuerza la interpretación de que cualquier tratamiento de datos biométricos debe ser riguroso y estar alineado con las mejores prácticas de protección de datos.

8.2. *Limitaciones*

Las principales limitaciones al uso del escaneo del iris para identificación incluyen la necesidad de cumplir con regulaciones estrictas de protección de datos, garantizar el consentimiento informado de los individuos y proporcionar medidas de seguridad adecuadas para proteger los datos personales. La legislación española, reforzada por la STC 119/2022, subraya la importancia de estos requisitos para asegurar que los datos biométricos se manejen de manera ética y legal. Además, cualquier uso de esta tecnología debe considerar el contexto específico y las regulaciones aplicables, como las que se aplican en contextos laborales o en la Comunidad Valenciana para los establecimientos de juego. Por ejemplo, en el ámbito laboral, es crucial que el uso del escaneo de iris cumpla con las regulaciones sobre protección de datos y derechos laborales, garantizando que no se vulneren los derechos de los empleados.

Aunque el escaneo de iris es legalmente permitido en contextos específicos de aplicación de la ley, su uso podría estar sujeto a restricciones en otros contextos, especialmente en relación con la protección de datos personales y la privacidad individual. La legislación sobre protección de datos personales puede imponer limitaciones en el tratamiento de datos biométricos, como el escaneo de iris, fuera de los contextos específicamente amparados por la ley. Por ejemplo, fuera del ámbito de la seguridad y la aplicación de la ley, cualquier uso de tecnologías de identificación biométrica debe cumplir con los principios de proporcionalidad, necesidad y minimización de datos, garantizando que solo se recopilen y procesen los datos estrictamente necesarios para el propósito específico.

Además, el uso de tecnologías biométricas debe estar alineado con los derechos fundamentales de los individuos, asegurando que

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

cualquier implementación respete la privacidad y no conduzca a un uso excesivo o indebido de los datos personales. Las organizaciones deben ser transparentes en sus prácticas de manejo de datos biométricos y proporcionar a los individuos información clara sobre cómo se utilizarán sus datos, las medidas de seguridad implementadas y los derechos que tienen para controlar el uso de su información personal. Estas limitaciones y requisitos reflejan la necesidad de un enfoque cuidadoso y bien regulado en el uso del escaneo de iris, asegurando que esta tecnología se utilice de manera responsable y en consonancia con las normativas vigentes, protegiendo así los derechos y la privacidad de los individuos en un entorno cada vez más digitalizado.

8.3. *Fundamentos del escaneo de iris*

El escaneo de iris es una tecnología biométrica que utiliza patrones únicos en el iris del ojo para identificar a las personas. El iris es la parte coloreada del ojo que rodea la pupila y tiene patrones complejos que son únicos para cada individuo. Estos patrones se forman en los primeros años de vida y permanecen prácticamente inalterados a lo largo del tiempo, lo que los convierte en una característica ideal para la identificación biométrica. El proceso de escaneo de iris implica capturar una imagen de alta resolución del iris utilizando una cámara especializada. Esta imagen se procesa mediante algoritmos que extraen características específicas del patrón del iris, creando una plantilla biométrica. Esta plantilla se almacena en una base de datos y se utiliza para comparar con futuras imágenes del iris para verificar la identidad de la persona.

El escaneo de iris tiene una amplia gama de aplicaciones debido a su alta precisión y fiabilidad. En el sector de la seguridad, se utiliza en aeropuertos y fronteras para verificar la identidad de los viajeros y prevenir el uso de documentos falsificados. También se emplea en instalaciones de alta seguridad, como centrales nucleares y edificios gubernamentales, para controlar el acceso y garantizar que solo las personas autorizadas puedan entrar. Asimismo, en el ámbito financiero, algunas instituciones utilizan el escaneo de iris para autenticar a los clientes en cajeros automáticos y servicios bancarios en línea, proporcionando una capa adicional de seguridad. Además, en dispositivos móviles, el escaneo de iris se utiliza como una forma segura de desbloquear teléfonos y acceder a aplicaciones sensibles.

8.4. *Implicaciones de privacidad*

La recopilación y el almacenamiento de datos biométricos del iris plantean importantes cuestiones de privacidad. Dado que estos datos son extremadamente sensibles y únicos para cada individuo, su manejo requiere un cuidado especial. Las organizaciones deben asegurar que la recopilación de datos biométricos se realice con el consentimiento informado de los individuos y que estos datos se almacenen de manera segura para evitar accesos no autorizados y posibles filtraciones (Martínez Martínez 2020).

A pesar de la robustez del escaneo de iris, existen riesgos de seguridad y posibles vulnerabilidades. Los sistemas de escaneo de iris pueden ser objeto de ataques, como la falsificación de plantillas biométricas o la captura de imágenes de alta resolución sin el conocimiento del individuo. Es crucial implementar medidas de seguridad avanzadas, como el cifrado de datos biométricos y la detección de *liveness* (pruebas de vida), para mitigar estos riesgos y proteger la integridad de los datos biométricos.

El uso indebido de datos biométricos del iris puede tener graves consecuencias para la privacidad de los individuos. Si estos datos caen en manos equivocadas, pueden ser utilizados para el robo de identidad, seguimiento no autorizado y otras actividades malintencionadas. Además, la recopilación masiva de datos biométricos puede llevar a la creación de bases de datos que, si no se gestionan adecuadamente, pueden ser vulnerables a brechas de seguridad. Por este motivo, es fundamental que las organizaciones que utilicen escaneo de iris adopten políticas claras y transparentes sobre el manejo de datos biométricos, aseguren que se cumplan las normativas de protección de datos y proporcionen a los individuos información y control sobre el uso de sus datos personales. Esto ayudará a proteger la privacidad y los derechos fundamentales de los individuos en una era de creciente digitalización y dependencia de tecnologías biométricas.

8.5. *Desafíos éticos y legales*

Uno de los principales desafíos éticos en el uso del escaneo de iris es garantizar que los individuos otorguen su consentimiento informado de manera libre y voluntaria. Esto significa que las personas deben estar plenamente informadas sobre cómo se recopilarán, utilizarán y almacenarán sus datos biométricos, así como sobre sus derechos para acceder, rectificar y eliminar esta información. La

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

autonomía del individuo debe ser respetada, permitiéndole tomar decisiones informadas sobre su participación en sistemas que utilizan escaneo de iris.

El tratamiento de datos biométricos está sujeto a estrictas normativas y regulaciones diseñadas para proteger la privacidad de los individuos. En la Unión Europea, el RGPD establece principios claros sobre el procesamiento de datos biométricos, incluyendo la necesidad de bases legales sólidas, como el consentimiento explícito o intereses legítimos claramente justificados. Además, la Ley Orgánica 3/2018 en España refuerza estas protecciones a nivel nacional, asegurando que cualquier uso de datos biométricos cumpla con los estándares legales más rigurosos.

Las organizaciones que recopilan y procesan datos biométricos del iris tienen la responsabilidad de implementar medidas adecuadas para proteger estos datos. Esto incluye la adopción de políticas de privacidad robustas, la formación de empleados en prácticas seguras de manejo de datos y la implementación de tecnologías avanzadas para prevenir accesos no autorizados. Las organizaciones deben ser transparentes en sus prácticas y ser responsables ante cualquier violación de privacidad o mal manejo de los datos biométricos.

8.6. *Estrategias de mitigación de riesgos*

Para minimizar el riesgo de violaciones de privacidad, las organizaciones deben adoptar un enfoque proactivo en la gestión de datos biométricos del iris. Esto incluye la realización de evaluaciones de impacto sobre la privacidad antes de implementar tecnologías de escaneo de iris, así como la revisión regular de las prácticas de manejo de datos para identificar y mitigar posibles riesgos.

Las buenas prácticas en la gestión de datos biométricos del iris incluyen la limitación de la recopilación de datos al mínimo necesario, el almacenamiento seguro de estos datos y la implementación de políticas claras sobre su uso y retención. Además, es crucial asegurar que los datos sean accesibles solo a personal autorizado y que se mantengan registros detallados de cualquier acceso o uso de los datos biométricos. Además, el uso de tecnologías de cifrado y anonimización es fundamental para proteger la privacidad de los usuarios cuyos datos biométricos se recopilan. El cifrado asegura que los datos sean ininteligibles para cualquier persona no autorizada que pueda acceder a ellos, mientras que la anonimización reduce el riesgo de que los datos puedan ser asociados con individuos específicos. Estas

tecnologías deben ser parte integral de cualquier sistema que maneje datos biométricos del iris.

8.7. *Consideraciones éticas y sociales*

Encontrar un equilibrio entre la necesidad de seguridad y la protección de la privacidad es un desafío constante en el uso del escaneo de iris. Si bien esta tecnología ofrece altos niveles de seguridad y precisión en la identificación, su implementación debe ser cuidadosamente gestionada para evitar invasiones de privacidad y garantizar que los derechos individuales no sean comprometidos.

El uso del escaneo de iris puede afectar la percepción de privacidad personal de los individuos. La sensación de estar constantemente vigilado y la intrusión en la vida personal pueden generar desconfianza y resistencia hacia esta tecnología. Es crucial abordar estas preocupaciones mediante la transparencia, la educación y la comunicación abierta sobre los beneficios y las medidas de protección asociadas con el escaneo de iris.

La utilización de datos biométricos del iris para fines diversos, más allá de la seguridad y la autenticación, plantea importantes cuestiones éticas. Esto incluye el uso de datos biométricos en el marketing, la investigación y otros ámbitos que pueden no estar claramente justificados desde una perspectiva de privacidad y derechos individuales. Las organizaciones deben considerar cuidadosamente los impactos éticos de sus prácticas y asegurar que cualquier uso de datos biométricos esté alineado con los valores y principios éticos aceptados.

8.8. *El proyecto Worldcoin*

La AEPD emitió una medida cautelar que prohíbe a Worldcoin continuar recopilando y tratando los datos biométricos del iris de los usuarios en España[15]. Esta medida implica que la empresa no puede

[15] Ref.: EXP202312448 Asunto: Acuerdo de adopción de medida provisional. Véase: https://www.aepd.es/documento/co-000297-2023-medida-provisional.pdf. La Sección Primera de la Sala de lo Contencioso – Administrativo de la Audiencia Nacional avaló, a través de auto dictado el 11 de marzo de 2024, el Acuerdo de la AEPD. Véase: https://www.poderjudicial.es/cgpj/es/Poder-Judicial/Noticias-Judiciales/La-Audiencia-Nacional-avala-el-cese-cautelar-de-la-recopilacion-de-datos-a-traves-del-iris-de-Worldcoin-acordado-por-la-Agencia-de-Proteccion-de-datos.

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

seguir operando en centros comerciales situados en territorio español, en los que se ofrecía intercambiar información del iris a cambio de criptomonedas WLD. La AEPD destaca en su acuerdo que el iris es un dato personal y biométrico que requiere una protección especial, señalando además que Worldcoin vulnera la normativa europea al no facilitar información adecuada a los usuarios, ni permitirles retirar su consentimiento o eliminar la información recopilada.

Worldcoin no exigía documentos de identidad a los usuarios y se comprometió a llevar a cabo únicamente el escaneo de mayores de edad, aunque ha quedado acreditado que también escaneó el iris de menores de edad. La medida cautelar de la AEPD puede extenderse por hasta tres meses, y se comparte con el Comité Europeo de Protección de Datos, dado que Worldcoin opera a nivel comunitario. La empresa enfrenta posibles multas de hasta el 4% de sus ingresos anuales si no cumple con la orden de la AEPD. En respuesta, el jefe de protección de datos de Worldcoin ha criticado la medida, acusando a la AEPD de difundir afirmaciones inexactas y eludir la ley de la UE.

El asunto que nos ocupa presenta varios aspectos relevantes en términos de protección de datos y cumplimiento normativo. Entre estos, podemos destacar los siguientes:

a) Protección de datos personales: La recopilación y tratamiento de datos biométricos, como el escaneo del iris, se considera una actividad especialmente sensible desde el punto de vista de la protección de datos personales. La normativa europea, en particular el RGPD, establece requisitos estrictos para el manejo de estos datos, incluyendo la necesidad de obtener un consentimiento informado y garantizar la seguridad y privacidad de la información.
b) Consentimiento informado y autonomía del individuo: Uno de los principales puntos de conflicto en el caso de Worldcoin es la falta de un consentimiento informado adecuado por parte de los usuarios. La ley exige que el consentimiento para el tratamiento de datos biométricos sea específico, informado y otorgado libremente. En este caso, la empresa ha sido acusada de no proporcionar información suficiente sobre el proceso de escaneo del iris y de no permitir a los usuarios retirar su consentimiento.
c) Derechos de los individuos y responsabilidad de las organizaciones: La normativa de protección de datos otorga importantes derechos a los individuos, incluido el derecho a la supresión de datos y a la privacidad. Las empresas, como Worldcoin, tienen la ineludible responsabilidad de respetar estos derechos y garantizar que el tratamiento de datos se realice de

manera legal y ética. La infracción de estas normas puede resultar en sanciones significativas, incluyendo multas.

Conclusiones y recomendaciones

Primera. La aplicación de la IA en el tratamiento de datos biométricos presenta desafíos éticos considerables, desde sesgos y discriminación hasta preocupaciones sobre la privacidad y la seguridad de los datos. Estos desafíos tienen implicaciones directas en los derechos fundamentales de los individuos, exigiendo una evaluación y abordaje ético y jurídico integral.

Segunda. La falta de transparencia en los algoritmos de IA y en la recopilación de datos biométricos plantea inquietudes significativas. Garantizar la transparencia y la explicabilidad en el desarrollo y aplicación de sistemas de IA es esencial para fomentar la confianza, facilitar la rendición de cuentas y permitir que los individuos comprendan cómo se toman las decisiones que los afectan.

Tercera. La ausencia de un marco regulatorio sólido contribuye a la incertidumbre ética y jurídica en el uso de la IA y datos biométricos. Se requiere una legislación clara y específica que aborde los desafíos identificados, establezca límites éticos y defina responsabilidades para garantizar un despliegue ético y respetuoso de estas tecnologías. Las investigaciones indican que es esencial desarrollar marcos regulatorios robustos y éticamente responsables para el uso de tecnologías biométricas, como se ha argumentado en varios estudios recientes (Garrós 2021). Esto garantizará no solo la protección de los derechos individuales sino también el uso seguro y eficaz de estas tecnologías en la sociedad moderna.

Cuarta. La preservación de derechos fundamentales, como la igualdad, la privacidad o la interdicción de la discriminación, debe ser prioritaria en el desarrollo y aplicación de sistemas de IA. Los sesgos y discriminación inherentes deben abordarse para evitar violaciones de derechos fundamentales y garantizar que estos sistemas beneficien a la sociedad en su conjunto.

Quinta. La evolución rápida de la tecnología exige una consideración ética continua en el desarrollo e implementación de sistemas de IA. Los diseñadores, desarrolladores y responsables de políticas deben comprometerse a evaluar y mitigar los impactos éticos a medida que surgen nuevas aplicaciones y desafíos.

Sexta. Debemos establecer políticas claras y transparentes sobre el uso de datos biométricos del iris, asegurando que los individuos

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 341-374
https://doi.org/10.18543/djhr.3199 • http://djhr.revistas.deusto.es/

comprendan cómo se recopilarán, utilizarán y protegerán sus datos. Obtener consentimiento informado de manera clara y comprensible, respetando la autonomía del individuo y sus derechos sobre sus datos personales. Garantizar el cumplimiento de normativas como el RGPD y la Ley Orgánica 3/2018, que establecen estándares estrictos para el tratamiento de datos biométricos y la protección de la privacidad.

Séptima. Implementar medidas de seguridad robustas, incluyendo tecnologías avanzadas de cifrado y anonimización, para proteger los datos biométricos del iris contra accesos no autorizados y posibles vulnerabilidades es una necesidad imperiosa en estos momentos. Además, debemos proporcionar formación continua a empleados y usuarios sobre las mejores prácticas en el manejo de datos biométricos, concienciándolos sobre la importancia de la privacidad y sus derechos respecto a sus datos personales.

Octava. Urge investigar y desarrollar técnicas avanzadas para mejorar la anonimización y el cifrado de datos biométricos, garantizando una protección efectiva de la privacidad. Evaluar el impacto social y ético del escaneo de iris en diferentes contextos, identificando posibles riesgos y beneficios, así como adaptar normativas para abordar los avances tecnológicos y los desafíos emergentes en la protección de datos biométricos.

Novena. Por último, subrayar que debemos fomentar una mayor colaboración interdisciplinaria entre expertos en tecnología, ética, derecho y políticas públicas. Esto permitirá anticipar mejor los riesgos emergentes en el uso de datos biométricos y la IA, y crear marcos adaptativos que se mantengan actualizados con los rápidos avances tecnológicos. Además, se deberían promover iniciativas educativas para el público en general, aumentando la conciencia sobre los derechos de privacidad y el impacto que las tecnologías biométricas pueden tener en su vida cotidiana, garantizando una adopción más responsable y ética a nivel global.

Podemos afirmar que el uso de la IA en el tratamiento de datos biométricos presenta oportunidades significativas, pero también plantea riesgos sustanciales para los derechos fundamentales y la equidad. Abordar estos desafíos requiere una colaboración integral entre expertos en ética, juristas, tecnólogos y responsables de políticas para desarrollar soluciones que equilibren la innovación con la protección de los derechos individuales y colectivos. Es imperativo que futuras investigaciones continúen explorando los impactos legales y sociales del uso de datos biométricos, con énfasis en la necesidad de regulaciones más claras y específicas. Además, es recomendable que se realicen estudios empíricos para evaluar la efectividad de las políticas actuales y su implementación práctica.

Bibliografía

Aliaga, Laura, y Mª Estrella Gutiérrez. 2020. «Reino Unido: Reconocimiento facial en lugares públicos realizado por Fuerzas y Cuerpos de Seguridad en el marco de la prevención e investigación de delitos. La visión del ICO». *La Ley privacidad* 3: 18.

Álvarez, Nelia, y David Sanz. 2021. «Advertencia de la Agencia Española de Protección de Datos con respecto a la utilización de técnicas de e-proctoring en la evaluación online por la UNIR». *La Ley privacidad* 10: 6.

Andúgar, Miguel Á. 2023. «Videovigilancia, control de accesos y datos biométricos». En *La protección de datos en el ámbito parlamentario: guía práctica*, editado por Esther de Alba, 201-219. Madrid: Asociación de Delegados y Delegadas de Protección de Datos de Parlamentos.

Arroyo, Alicia. 2022. «Vehículos autónomos, responsabilidad y seguro: Avances legislativos y perspectivas». *Revista de Derecho del Sistema Financiero: mercados, operadores y contratos* 3: 4.

Baz, Jesús. 2021. *Los nuevos derechos digitales laborales de las personas trabajadoras en España: vigilancia tecnificada, teletrabajo, inteligencia artificial, Big Data*. Madrid: Wolters Kluwer España.

Barona, Silvia. 2024. «Tecnología biométrica y datos biométricos. Bondades y peligros. No todo vale». *Actualidad Jurídica Iberoamericana* 21: 298-331.

Bestard, Juan J. 2021. *La gestión de datos personales y el delegado de protección de datos en la sanidad pública: Con atención especial a la Comunidad de Madrid*. Tesis doctoral. Acceso el 18 de octubre de 2024: https://repositorio.uam.es/handle/10486/699704

Castellanos, Jorge. 2023. *Inteligencia artificial y democracia: garantías, límites constitucionales y perspectiva ética ante la transformación digital*, Barcelona: Atelier.

Castellanos, Jorge. 2024. «Una reflexión acerca de la influencia de la inteligencia artificial en los derechos fundamentales». en *Ciencia de datos y perspectivas de la inteligencia artificial*, editado por Francisca Ramón, 271-300. Valencia: Tirant Lo Blanch.

Cotino, Lorenzo. 2022. «Sistemas de inteligencia artificial con reconocimiento facial y datos biométricos. Mejor regular bien que prohibir mal». *El Cronista del Estado Social y Democrático de Derecho* 100: 68-79. Acceso el 18 de octubre 2024. https://www.uv.es/cotino/publicaciones/cronistacotinopublicado.pdf

Cotino, Lorenzo. 2023. «Reconocimiento facial automatizado y sistemas de identificación biométrica bajo la regulación superpuesta de inteligencia artificial y protección de datos». *Derecho público de la inteligencia artificial*, 347-402. Acceso el 18 de octubre 2024: https://www.fundacionmgimenezabad.es/sites/default/files/Publicar/publicaciones/documentos/oc27_13_lorenzo_cotino_es_o.pdf

Cotino, Lorenzo. 2023. «Una regulación legal y de calidad para los análisis automatizados de datos o con inteligencia artificial. Los altos estándares

que exigen el Tribunal Constitucional alemán y otros tribunales, que no se cumplen ni de lejos en España», *Revista General de Derecho Administrativo* 63: Iustel (RI §425995)

Díaz, Juan. 2023. «Datos biométricos para el control de presencia y accesos». *I+S: Revista de la Sociedad Española de Informática y Salud* 157: 62. Acceso el 13 de julio de 2024. https://seis.es/is-157/

Díaz Lima, David. 2023. «Datos biométricos y acceso a eventos deportivos. Comentarios al Informe de la AEPD». *La Ley privacidad* 15.

Espuga, Gerard. 2023. «La (i)licitud del tratamiento de datos biométricos para el registro de jornada». en *Más allá de la oficina: desafíos laborales emergentes en un mundo hiperconectado*, editado por Francisco Trujillo, 139-160. Cizur Menor: Aranzadi.

Etxeberria, José F. Pilar Martín, César A. Villegas y José L. Rodríguez. 2023. «Datos biométricos y reconocimiento facial en el proceso penal». En *La tecnología y la inteligencia artificial al servicio del proceso*, dirigido por María Luis Domínguez, Pilar A. Villegas y José L. Rodríguez Lainz, 107-126. A Coruña: Colex.

Flórez, Mª Lorena, y Angélica Mª Camelo. 2023. «Tecnologías de reconocimiento facial en Colombia: Análisis comparativo en relación con la protección de datos». *Ius et Praxis* 29 (1): 3-26. Acceso el 14 de junio de 2024: https://www.revistaiep.utalca.cl/wp-content/uploads/2023/03/02.-Florez-Maria-Lorena-y-Camelo-Angelica.pdf

Garrido, Andrea. 2023. «El derecho al respeto a la vida privada: ¿el precio a pagar por una Europa segura en la era tecnológica?» *Revista Integración Regional & Derechos Humanos* 11 (2). Acceso el 1 de junio de 2024: http://www.derecho.uba.ar/institucional/centro-de-excelencia-jean-monnet/revista-electronica/009/garrido-raya.pdf

Garriga, Ana, Cristina Pauner, Rosario García Mahamut y Beatriz Tomás. 2023. «La especial posición de los datos biométricos en el RGPD: peculiaridades derivadas de su naturaleza y riesgos asociados a su tratamiento», en *La implementación del reglamento general de protección de datos en España y el impacto de sus cláusulas abiertas,* coordinado por Jorge A. Viguri, 115-144. Valencia: Tirant lo Blanch.

Garrós, Imma. 2021. «Las categorías especiales de datos personales y su régimen aplicable». *Revista Aranzadi Doctrinal* 2.

González Calvo, Marcos. 2022. «¿Hacia nuevas restricciones en el uso de datos biométricos?» *Actualidad jurídica Aranzadi* 990.

Han, Byung–Chul. 2020. «La emergencia viral y el mundo del mañana». *El País.* 22 de marzo. Acceso el 22 de mayo de 2024. https://elpais.com/ideas/2020-03-21/la-emergencia-viral-y-el-mundo-de-manana-byung-chul-han-el-filosofo-surcoreano-que-piensa-desde-berlin.html.

Harari, Yuval N. 2024. *Nexus. Una breve historia de las redes de información desde la Edad de Piedra hasta la IA*. Barcelona: Debate.

Iturmendi, José M. 2023. «La discriminación algorítmica y su impacto en la dignidad de la persona y los derechos humanos: Especial referencia a los

inmigrantes». *Revista Deusto de derechos humanos* 12: 257-284. https://djhr.revistas.deusto.es/article/view/2910. Acceso el 2 de mayo de 2024

Martínez Martínez, Ricard. 2020. «Tecnología de verificación de identidad y control en exámenes online». *Revista de educación y derecho* 22. doi.org/10.1344/REYD2020.22.32357

Ponce, Julio. 2024. «Inteligencia Artificial. Decisiones administrativas discrecionales totalmente automatizadas y alcance del control judicial: ¿Indiferencia, insuficiencia o deferencia?», *Revista de Derecho Público: Teoría y Método* 9: 172-220. DOI: 10.37417/RPD/vol_9_2024_2151

Razquin, Martín M. 2022. «La identidad digital como derecho». *Derecho Digital e Innovación* 14.

Romano, Andrea. 2023. «Derechos fundamentales e inteligencia artificial emocional en iBorderCtrl: retos de la automatización en el ámbito migratorio». *Revista catalana de dret públic* 66: 237-252. doi.org/10.58992/rcdp.i66.2023.3928.

Sebé, Sonia, y Manel Santilari. 2022. «Los datos biométricos como categorías especiales de datos. Debate a raíz de las directrices del Comité Europeo de Protección de Datos sobre reconocimiento facial». *Comunicaciones en propiedad industrial y derecho de la competencia* 97: 23-43.

Sempere, Javier. 2020. «¿Se puede utilizar la huella para el control de accesos a un gimnasio?». *La Ley privacidad* 3.

Serrano, Fernando. 2022. *El uso de la inteligencia artificial para optimizar los ingresos tributarios.* Informe 7. Caracas: CAF. https://scioteca.caf.com/handle/123456789/1933 Acceso el 18 de marzo de 2024.

Simón, Pere. 2021. *Justicia cautelar e inteligencia artificial. La alternativa a los atávicos heurísticos artificiales*. Barcelona: Bosch Editores.

Simón, Pere y Lorenzo Cotino (Dirs.). 2024. *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*. Madrid: Aranzadi.

II

Book reviews

Críticas bibliográficas

Balcerzak, Michal and Julia Kapelańska-Pręgowska, eds. 2024. *Artificial Intelligence and International Human Rights Law. Developing Standards for a Changing World*. Cheltenham: Edward Elgar. 347 p.

doi: https://doi.org/10.18543/djhr.3200

E-published: December 2024

The book *Artificial Intelligence and International Human Rights Law*, edited by Michał Balcerzak and Julia Kapelańska-Pręgowska, offers a meticulous examination of the interplay between human rights and artificial intelligence (AI) within the framework of international law. The written composition is divided into two parts. The introductory part presents a thorough analysis of the in-force global regulatory frameworks. The final part underscores the importance of conducting specific targeted sectoral evaluations to meticulously examine the ramifications of artificial intelligence across essential domains such as justice, privacy, health, and commerce. The authors and contributors analyze, in detail, the ethical, legal, and political quandaries that artificial intelligence presents across its seventeen chapters. The aim is to formulate strategies that guarantee such technology upholds fundamental rights instead of compromising them. This book's significant contribution to AI ethics and human rights is evident in its comprehensive analysis and the strategies it proposes to ensure AI respects and promotes human rights. Although this purpose aims for a commendable level of effort and ambition, it also reveals weaknesses in integration and coherence that prevent a unified narrative. Nevertheless, their contributions show a notable intellectual rigor and a capacity to illuminate pressing matters within global technological governance.

The book begins by asserting the significant role of global entities, including the United Nations (UN), the Council of Europe, and the European Union (EU), in the governance of artificial intelligence. The first chapter, authored by Michał Balcerzak, focuses on the implications of UN human rights standards in AI governance. Balcerzak critically examines the efforts undertaken by the UN system from 2019 to 2023, highlighting key documents such as the Governing AI for Humanity report, which advocates for ethical and transparent regulation. He emphasizes how existing standards, such as the Guiding Principles on Business and Human Rights, provide a solid normative foundation but

remain insufficient to address challenges like algorithmic bias, mass surveillance, and structural discrimination. This chapter articulates the shortcomings of traditional regulatory frameworks while proposing a combination of ethical principles and practical tools, such as UNESCO's ethical impact assessments, challenging the reader's perspective and understanding of AI governance.

In the second chapter, Elżbieta Hanna Morawska's examination of the Council of Europe's normative advances in AI regulation, particularly the work of the Ad Hoc Committee on Artificial Intelligence (CAHAI) and its evolution into the Committee on Artificial Intelligence (CAI), has practical implications. Her detailed analysis of the fundamental principles underlying the developing Framework Convention, such as transparency, accountability, and data protection, and her call to overcome regulatory fragmentation and ensure non-state actors' inclusion in policymaking provide actionable insights for policymakers and practitioners.

The third chapter, written by the thorough Piotr Staszczyk, is notable for its length and takes an in-depth look at the EU. The Digital Services Regulation 2022/2065, which entered into force on February 17, 2024, has established a more comprehensive framework that classifies AI applications by risk levels and imposes strict requirements for high-impact technologies such as real-time facial recognition. The chapter combines technical analysis with a comparative perspective, exploring tensions between EU regulation and more permissive approaches in the United States and China. The EU's regulatory proposal is presented as a global model for balancing technological innovation with protecting fundamental rights.

In the fourth chapter, Marya Akhtar and Rikke Frank Jørgensen address the impact of automated decision-making systems (ADM) in the public sector. Using practical cases from Denmark, the authors illustrate how ADMs can enhance administrative efficiency, perpetuate discrimination, and infringe on fundamental legal principles. This chapter is precious because it emphasizes algorithmic transparency and ethical impact assessments as essential tools to mitigate ADM-associated risks.

In the fifth chapter, Agnieszka Bień-Kacała examines the impact of Pegasus spyware on human rights and democratic processes, emphasizing its misuse in countries such as Poland and Hungary to surveil political opponents and journalists. Bień-Kacała critiques the inadequacy of existing regulatory frameworks to prevent abuses linked to mass surveillance technologies, proposing measures such as strict prohibitions on their use in politically sensitive contexts. This chapter's

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 377-384
https://doi.org/10.18543/djhr.3200 • http://djhr.revistas.deusto.es/

critique of existing regulatory frameworks is a significant contribution to the scholarly debate on AI governance and human rights protection. The chapter combines rigorous empirical analysis with concrete normative proposals, reinforcing its relevance in the contemporary context.

In the sixth chapter, authored by Julia Kapelańska-Pręgowska, Emilia Sarnacka, and Katarzyna Syroka-Marczewska, the interdisciplinary nature of their analysis is highlighted. Their examination of the implications of AI in healthcare, which emphasizes its benefits and risks, combines legal, ethical, and practical perspectives. This comprehensive approach and their emphasis on human oversight as an essential principle for AI governance in health make their chapter stand out. The book's emphasis on human oversight as a practical recommendation for AI governance in health is a significant contribution to the field, as it provides a clear path for policymakers and practitioners to follow.

In the seventh chapter, Joanna Mazur and Zuzanna Choińska analyze the tensions between public security and fundamental rights protection using facial recognition technologies. The authors identify deficiencies in implementing European regulations through case studies such as Clearview AI and advocate for stricter oversight and transparency measures. This chapter is a notable contribution to the surveillance and human rights debate.

In the eighth chapter, Ewa Michałkiewicz-Kądziela examines deepfakes as a growing threat to human rights and democracy. The chapter highlights how these technologies, used to manipulate audiovisual content, can violate privacy, damage reputations, and undermine trust in democratic processes. Michałkiewicz-Kądziela advocates for the global regulation of deepfakes, complemented by educational strategies to strengthen digital literacy.

The ninth chapter, by Ewa Milczarek, addresses the legal challenges posed by AI-generated creativity. It questions traditional concepts of authorship and intellectual property. It proposes innovative solutions, such as contractual models and specific licenses, to ensure that machine-generated works are regulated relatively and effectively. This chapter is a crucial contribution to the debate on intellectual property in the age of AI.

In the tenth chapter, Anne Oloo explores the challenges and opportunities posed by algorithmic media in Africa, analyzing their impact on human rights, democracy, and social structures. The author highlights how AI systems, used to personalize content or combat disinformation, can also exacerbate structural inequalities due to

algorithmic biases and a lack of representativeness in data. A key example is content moderation tools, which often misinterpret or overlook cultural and linguistic nuances, leading to digital exclusion. The author underscores the importance of initiatives like the African Union's Blueprint for AI in Africa, which seeks to foster regional collaboration and harmonize national data protection laws. This holistic approach reflects a commitment to integrating regional particularities into global governance debates, making this chapter essential to understanding how local contexts shape AI implementation and regulation.

In the eleventh chapter, Maria O'Sullivan analyzes the difficulties traditional legal systems face in providing effective remedies to victims of human rights violations caused by AI. The author identifies issues such as algorithmic opacity, which makes it difficult for victims to understand and challenge automated decisions, and the lack of legal mechanisms to address systemic violations created by these technologies. Examples like the SyRI system in the Netherlands or Australia's Robodebt program illustrate how automated tools can perpetuate structural inequalities and infringe on fundamental rights such as privacy and equality before the law when designed without adequate safeguards. O'Sullivan advocates for a comprehensive approach to redress, combining collective measures such as group claims with mandatory ethical impact assessments and greater transparency in system design. This chapter provides a critical perspective on the limitations of existing regulatory frameworks. It emphasizes adapting legal systems to ensure human rights are not compromised in the digital era.

The twelfth chapter, by Joanna Rezmer, focuses on how AI is transforming the labor sector, highlighting its benefits and risks. Rezmer explores how automated technologies reshape recruitment, performance evaluation, and resource management processes, raising serious concerns about worker privacy, decision-making transparency, and perpetuating discriminatory biases. For instance, algorithmic recruitment tools have been shown to reinforce historical patterns of exclusion, disproportionately affecting women and minority groups. The author also examines the risks of job precarity and increasing inequality, noting that while automation can enhance workplace productivity and safety, it can also concentrate economic benefits in the hands of a few. Rezmer calls for adopting international standards to regulate the implementation of AI in the workplace, emphasizing the role of the International Labour Organization in promoting a human-centered approach. This chapter combines rigorous empirical

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 377-384
https://doi.org/10.18543/djhr.3200 • http://djhr.revistas.deusto.es/

analysis with concrete normative proposals, making it a critical contribution to the debate on the future of work.

In the thirteenth chapter, Maciej Jerzy Siwicki analyzes the impact of scalper bots in e-commerce, a phenomenon that has gained prominence in the digital economy. Scalper bots are designed to purchase high-demand products and resell them at inflated prices, creating artificial scarcity and harming consumers and businesses. The author examines specific cases, such as mass purchases of event tickets or gaming consoles, to illustrate how these practices negatively impact market fairness and consumer trust. From a regulatory perspective, Siwicki discusses the European legal framework, including the GDPR and the proposed AI Act, highlighting how these instruments aim to regulate bot usage and protect consumer rights. A notable strength of the chapter is its critical analysis of the limitations of existing technical and legal measures and its proposal for a combined approach integrating regulatory solutions, advanced technological tools, and public awareness campaigns.

In the fourteenth chapter, Tomasz Sroka examines how AI is transforming judicial administration, highlighting the benefits this technology offers in terms of efficiency and warning of the risks it poses to fundamental guarantees of the right to a fair trial. Sroka emphasizes that while AI can facilitate administrative tasks and improve predictability in judicial processes, its use in critical decisions raises serious challenges regarding transparency, accountability, and human oversight. Algorithmic opacity, known as the "black box effect", makes it difficult for involved parties to understand and challenge automated decisions, threatening fundamental principles such as equality of arms and the right to adequate defense. The chapter proposes normative safeguards such as mandatory human oversight and algorithmic transparency, stressing that judicial decisions must permanently preserve the central role of human judges. This chapter is particularly relevant in increasing automation, offering a critical reflection on balancing technological innovation with fundamental rights.

In the fifteenth chapter, Agnieszka Szpak addresses the ethical and legal dilemmas associated with lethal autonomous weapons systems, highlighting their growing relevance in international humanitarian law debates. Szpak examines how these technologies, capable of selecting and attacking targets without direct human intervention, pose significant challenges to principles such as the distinction between combatants and civilians, proportionality, and attack precautions. The author discusses debates within the Convention on Certain

Conventional Weapons framework, emphasizing "meaningful human control" as an essential safeguard for ensuring accountability and legality in using these weapons. This chapter combines technical, ethical, and normative analysis, offering a comprehensive view of a topic of growing importance in international security.

In the sixteenth chapter, Lutiana Valadares Fernandes Barbosa and Ana Luísa Zago de Moraes analyze how AI is used in migration processes, highlighting its potential to increase efficiency and the ethical and legal risks it poses. The authors emphasize that using automated systems to determine refugee status can compromise fundamental principles such as due process and the principle of non-refoulement. This chapter stands out for its interdisciplinary approach, combining concrete examples with a solid normative framework, and its proposal of hybrid mechanisms integrating human and technological oversight to ensure fair and transparent decisions.

The seventeenth chapter, authored by Peng Wang and Guannan Qu, explores the implementation of Smart Courts in China. This model has revolutionized judicial administration through AI and big data. The authors describe how this transformation, divided into two main phases —process digitization and the integration of advanced technologies— has enhanced efficiency and transparency in judicial case handling. However, they also highlight these innovations' ethical and regulatory challenges, such as algorithmic opacity, threats to citizens' privacy, and the absence of adequate regulations to standardize application. Despite these challenges, the chapter underscores that while AI provides invaluable support in streamlining judicial systems, it cannot replace human oversight, which is essential to ensuring fairness and public trust. The authors advocate for clear safeguards to balance technological innovation with protecting fundamental rights.

This volume makes a unique and significant contribution to the discourse on AI governance and human rights. It presents a comprehensive research methodology and ethical goals that are commendable. The authors delve into the main challenges that artificial intelligence poses for human rights, covering privacy, justice, employment, and health. However, their reliance on established legal structures and a prevailing normative framework may limit their ability to provide innovative and adaptable solutions. The global landscape is rapidly transforming across numerous industries, driven by advances in artificial intelligence. These developments underscore the inherent deficiencies of international institutions and regulatory frameworks, which need to be revised to effectively address power imbalances and

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 377-384
https://doi.org/10.18543/djhr.3200 • http://djhr.revistas.deusto.es/

structural inequities exacerbated by the rise of emerging technologies. The book's failure to offer a comprehensive and inclusive point of view is a significant shortcoming. While it does make superficial allusions to the disparities in technological access, especially in regions such as Africa, it does not comprehensively examine the possibilities for improving these trends. The recognition of pioneering applications of artificial intelligence is evident in fields such as smart agriculture across the African continent. However, we still overlook the ramifications of relying on technologies conceived by external entities. The lack of this point of view reinforces a technocratic narrative that emphasizes technical solutions, overlooking the fundamental structural changes needed to achieve social and technological justice. In addition, the text does not define a feminist perspective sufficiently. While references to the various effects of artificial intelligence on women and marginalized groups are present, they appear fragmented and need a comprehensive, systematic, critical examination. The exclusion of gendered implications and fundamental power dynamics in the discourse around deepfakes and mass surveillance requires scrutiny, as these issues are frequently treated as ancillary rather than central to the analytical framework. The intersectional approach reveals a remarkable mismatch between the prevailing normative narratives expressed in the discourse and the authentic lived experiences of the most marginalized communities.

Nevertheless, this text is fundamental to understanding the intricate relationship between artificial intelligence and human rights. It not only highlights the limitations set by conventional normative approaches but also challenges them, shaping future discussions in the field. The text institutes the international institutions and regulatory frameworks within a historical context characterized by a gradual, rather than an exponential, evolution. In juxtaposition, the emergence of artificial intelligence is simultaneously transforming regulatory structures, leading legislators to get caught up in extensive deliberations that produce minimal substantive results. In light of the rapid advances within technological spheres, it is imperative to conduct a comprehensive analysis of the methodologies used in regulating and implementing artificial intelligence.

In conclusion, this volume significantly contributes to the discourse on AI governance and human rights. Its interdisciplinary approach, sectoral focus, and actionable recommendations provide a robust framework for addressing AI's ethical and legal challenges. The book's value lies in its ability to bridge theoretical principles with practical applications, paving the way for a more equitable and rights-centered

technological future. However, the book could benefit from greater cohesion and a more inclusive perspective, particularly regarding structural inequities and gendered impacts. Despite these limitations, it offers invaluable insights for academics, policymakers, and practitioners navigating the complexities of AI regulation.

Itziar Artíñano Ortiz
Universidad Complutense de Madrid

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 377-384
https://doi.org/10.18543/djhr.3200 • http://djhr.revistas.deusto.es/

Cotino, Lorenzo y Jorge Castellanos, eds. 2023. *Algoritmos abiertos y que no discriminen en el sector público*. Valencia: Tirant lo Blanch. 292 p.

doi: https://doi.org/10.18543/djhr.3201

Fecha de publicación en línea: diciembre de 2024

La inteligencia artificial (en adelante, IA) es un agente activador de procesos que afecta tanto a la vida cotidiana de las gentes, como a los procesos macrosociales de las colectividades. Así, la IA contribuye, en no pequeña medida, a la transformación del mundo. Por lo mismo, está transformando la administración pública, ofreciendo nuevas posibilidades para la gestión de datos, la prestación de servicios y la toma de decisiones. Sin embargo, esta transformación plantea desafíos que van más allá de lo técnico, ya que involucra cuestiones éticas, legales y sociales que tienen un impacto directo en los derechos fundamentales. Cuando los sistemas algorítmicos son utilizados para asignar recursos, evaluar beneficios o tomar decisiones automatizadas que afectan la vida de los ciudadanos, existe un riesgo real de vulnerar derechos como la igualdad ante la ley, la privacidad, el derecho a la participación política, el acceso a la justicia o la interdicción de la discriminación.

En este entorno resulta inevitable apuntar el riesgo que puede comportar la evolución de la tecnología inicialmente concebida como medio instrumental y aplicada, y que desde hace tiempo apunta una serie de preocupantes rasgos en orden a constituirse en un medio que puede socavar un buen número de derechos fundamentales. Así, la obra colectiva *Algoritmos abiertos y que no discriminen en el sector público*, coordinada por Lorenzo Cotino, Catedrático de Derecho Constitucional de la Facultad de Derecho de la Universidad de Valencia y por Jorge Castellanos, Profesor Titular de la misma asignatura en la misma Casa de Estudios, constituye, de manera indisputada, un aporte esencial que aborda de manera exhaustiva los aspectos nucleares de esta preocupación. Nos referimos a dos autores que desde hace ya unos años han contrastado su talento mostrado una inexhausta capacidad y una irrefrenable vocación en el abordaje de los aspectos tecnológicos y más estrictamente aplicados de la revolución digital y su intersección con el Derecho. Se configura en esta obra un conjunto que recoge colaboraciones, con un enfoque interdisciplinar, que combina reflexiones teóricas, análisis normativos y

estudios de casos y ofrece un marco integral para analizar y resolver los problemas éticos y legales que plantea la implementación de la IA en el sector público. Los coordinadores, así como la nómina de prestigiosos autores que colaboran en el volumen que aquí recensionamos, no solo identifican los riesgos asociados con la opacidad algorítmica, los sesgos en los datos y la automatización de decisiones, sino que también presentan soluciones concretas para garantizar que estos sistemas sean transparentes, responsables y respetuosos de los derechos fundamentales.

La relevancia de esta obra, que ofrece un tratamiento acabado de la materia, radica en su capacidad para vincular los desafíos tecnológicos con el marco de los derechos humanos. Los autores confirman su honda preocupación por los derechos fundamentales al afirmar que, sin una regulación adecuada y un diseño ético, los algoritmos pueden perpetuar desigualdades, reforzar sesgos históricos y socavar la confianza pública en las instituciones. Para evitar estas consecuencias, en el libro se subraya la necesidad de implementar mecanismos de rendición de cuentas y de evaluar previamente el impacto de los sistemas algorítmicos sobre las libertades individuales y colectivas. Además, aborda el reto de adaptar las normativas existentes para enfrentar los riesgos específicos que plantea la IA, explorando cómo los principios jurídicos clásicos –la proporcionalidad, la igualdad o la no discriminación– pueden aplicarse a este nuevo entorno tecnológico.

Uno de los capítulos centrales de la obra recensionada en estas líneas lleva por título "Las evaluaciones de impacto algorítmico en los derechos fundamentales: hacia una efectiva minimización de sesgos", escrito por Pere Simón. El profesor de la Universidad de Girona establece el marco conceptual para abordar el uso ético y responsable de la IA en el ámbito público. Simón introduce las evaluaciones de impacto algorítmico (EIA) como herramientas fundamentales para identificar, prevenir y mitigar los riesgos asociados al uso de algoritmos en decisiones administrativas que afectan derechos fundamentales. Tal y como pone de manifiesto Simón, estos sistemas, aunque se presentan como técnicamente avanzados, no están exentos de sesgos, ya que dependen de los datos utilizados para su entrenamiento y de las decisiones humanas implícitas en su diseño.

Simón realiza un análisis comparado de marcos normativos internacionales, destacando casos como el *Algorithmic Impact Assessment Tool* de Canadá, un modelo que obliga a las administraciones públicas a evaluar los riesgos de sus algoritmos antes de su implementación. También resalta el Reglamento de Inteligencia

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 385-392
https://doi.org/10.18543/djhr.3201 • http://djhr.revistas.deusto.es/

Artificial de la Unión Europea, que clasifica los sistemas de IA según su nivel de riesgo y establece requisitos específicos para aquellos considerados de alto riesgo. Estas normativas, explica Simón, no solo son aplicables a los contextos en los que fueron desarrolladas, sino que también pueden adaptarse a otros países, como España, donde todavía se carece de mecanismos estandarizados para supervisar y evaluar los algoritmos. El autor subraya que las EIA no deben concebirse como trámites burocráticos, sino como procesos dinámicos que se apliquen a lo largo de todo el ciclo de vida del sistema algorítmico. Propone que incluyan un análisis exhaustivo de los datos utilizados, las metodologías empleadas, los resultados esperados y las posibles repercusiones en los derechos de las personas afectadas. Además, destaca la importancia de incorporar la participación de múltiples actores, incluyendo auditores externos, expertos técnicos y comunidades afectadas, para garantizar una evaluación inclusiva y legítima. Este enfoque interdisciplinario, junto con ejemplos concretos de aplicación, hace que el capítulo de Simón sea un referente para abordar la problemática de los sesgos algorítmicos de manera práctica y efectiva.

El asunto de la discriminación algorítmica es asimismo abordado con detalle por Raquel Valle en el capítulo que lleva por nombre "Una inteligencia artificial a medida de las personas: el control de la discriminación algorítmica". Valle argumenta que, aunque los algoritmos tienen el potencial de mejorar la eficiencia y optimizar procesos, también pueden amplificar desigualdades existentes o generar nuevas formas de discriminación. Esto se debe principalmente a sesgos inherentes en los datos de entrenamiento o al diseño del sistema. La autora analiza casos concretos, como los sistemas de contratación automatizada que han descartado a candidatos pertenecientes a minorías o algoritmos financieros que han denegado créditos a determinados grupos sociales. Valle propone un enfoque integral para abordar la discriminación algorítmica, combinando soluciones técnicas, principios éticos y marcos normativos. Entre sus propuestas destaca la implementación de auditorías algorítmicas periódicas, que permitan evaluar y corregir sesgos de manera continua. También aboga por la supervisión humana en las decisiones críticas y por la adopción de criterios de equidad durante el diseño de los sistemas. Además, enfatiza que la gobernanza algorítmica debe incluir mecanismos claros de rendición de cuentas, que permitan a las personas afectadas cuestionar las decisiones tomadas por algoritmos y entender cómo fueron generadas.

Marco Emilio Sánchez Acevedo, abogado especializado en Derecho Digital, analiza en el capítulo "Buena administración algorítmica y

debido proceso frente a los sesgos", cómo integrar los sistemas algorítmicos en la administración pública sin comprometer principios fundamentales como la equidad y la transparencia. Sánchez identifica los principales tipos de sesgos algorítmicos, como el sesgo de representación, el sesgo de selección y el sesgo de optimización. Para abordar estos problemas, propone medidas como la supervisión humana, la publicación de evaluaciones de impacto y la realización de auditorías externas. Según el autor, estas medidas no solo garantizan la equidad en las decisiones, sino que también refuerzan la confianza ciudadana en las instituciones públicas.

La interoperabilidad de datos y la reutilización de tecnología son analizadas por Rubén Martínez Gutiérrez en el capítulo titulado "Datos abiertos, interoperabilidad y reutilización de tecnología para la inteligencia artificial del sector público". Martínez argumenta que los datos son la base de cualquier sistema algorítmico, pero su calidad y accesibilidad determinan la eficacia y la equidad de los resultados. Sin embargo, advierte que en el sector público persisten problemas como la fragmentación de los datos y la falta de estándares técnicos. Para superar estos desafíos, propone la creación de infraestructuras de datos interoperables y el uso de tecnologías de código abierto. También destaca la importancia de garantizar la privacidad y la seguridad de los datos, implementando medidas de protección como el cifrado y el acceso restringido.

Otro de los pilares de la obra es el análisis de Lorenzo Cotino relativo a los registros públicos de algoritmos, abordado en los capítulos "Cuándo deben crearse registros y dar transparencia a los algoritmos y sistemas de inteligencia artificial públicos" y "La información que hay que facilitar en los registros públicos de algoritmos y de inteligencia artificial". Capítulos en los que el profesor Cotino desarrolla una propuesta innovadora y práctica para implementar registros públicos como herramientas clave para garantizar la transparencia y la rendición de cuentas en el sector público. Nuestro autor enfatiza que la transparencia no puede ser uniforme, sino que debe ajustarse al nivel de impacto de los algoritmos en los derechos fundamentales. Por ejemplo, los sistemas que toman decisiones en áreas críticas, como la justicia o la sanidad, requieren estándares de transparencia mucho más rigurosos que aquellos utilizados en tareas técnicas de menor relevancia.

En el primer capítulo, Cotino establece un marco normativo claro para determinar cuándo y cómo deben implementarse los registros públicos. Argumenta que estos registros no solo tienen un valor técnico, sino también ético y político, ya que permiten a los ciudadanos

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 385-392
https://doi.org/10.18543/djhr.3201 • http://djhr.revistas.deusto.es/

supervisar y entender cómo las administraciones públicas utilizan la IA. Además, propone un modelo escalonado de transparencia que clasifica los algoritmos según su nivel de riesgo. Este enfoque asegura que los sistemas más complejos y con mayor impacto estén sujetos a evaluaciones más detalladas.

En el capítulo coescrito con Alba Soriano, profesora de Derecho Administrativo, los autores detallan la estructura y contenido que deberían tener estos registros y sugieren que los registros incluyan desde información básica sobre la finalidad del algoritmo hasta detalles técnicos como los datos de entrenamiento, las metodologías utilizadas y los resultados de las auditorías algorítmicas. Proponen un modelo de presentación por capas que garantice la accesibilidad de la información a distintos públicos: ciudadanos, expertos técnicos y responsables políticos. Este enfoque, inspirado en modelos de transparencia como el utilizado en la protección de datos personales, equilibra la necesidad de accesibilidad con la protección de la privacidad y la seguridad de los sistemas. La relevancia de estas propuestas no solo radica en su aplicabilidad práctica, sino también en su capacidad para fortalecer la confianza ciudadana en el uso de IA en el sector público. Al detallar cómo implementar registros públicos efectivos, Cotino y Soriano ofrecen una guía práctica que podría adoptarse en múltiples contextos administrativos.

En el capítulo "Gobernanza pública en materia algorítmica: una propuesta de formulación de los registros públicos", firmado por Jorge Castellanos y Adrián Palma, se trata la gobernanza como un elemento esencial para garantizar el uso ético y responsable de los algoritmos en el sector público. Los autores argumentan que la gobernanza algorítmica debe basarse en principios de equidad, transparencia y responsabilidad, y que estos principios deben guiar todas las etapas del ciclo de vida de un sistema algorítmico, desde su diseño hasta su implementación y evaluación continua. Castellanos y Palma proponen la creación de unidades administrativas especializadas dentro de las administraciones públicas, responsables de gestionar los registros de algoritmos, supervisar su uso y garantizar el cumplimiento de las normativas. Estas unidades tendrían la capacidad de clasificar los sistemas según su nivel de riesgo, realizar auditorías regulares y fomentar la participación ciudadana en la supervisión de los algoritmos. Además, los autores destacan la importancia de establecer marcos legales claros que definan las responsabilidades de las administraciones, los desarrolladores de algoritmos y otros actores involucrados.

Otro punto destacado del capítulo es la propuesta de Castellanos y Palma para fomentar la cooperación internacional en la gobernanza

algorítmica. Argumentan que los desafíos asociados al uso de IA en el sector público no pueden abordarse de manera aislada, ya que las tecnologías y los datos suelen trascender fronteras nacionales. Por ello, sugieren que los países compartan experiencias, buenas prácticas y estándares técnicos para garantizar un enfoque coherente y efectivo en la regulación de los algoritmos.

En los capítulos finales, el libro se adentra en el análisis de sectores específicos donde la IA tiene un impacto significativo. La iusadministrativista Rosa Cernada examina en "De la digitalización a la inteligencia artificial: el porvenir de la justicia en la Unión Europea", la capacidad transformativa de la IA en relación con el sistema judicial, mejorando su eficiencia y reduciendo los tiempos de resolución de casos. Sin embargo, Cernada también advierte sobre los riesgos éticos y legales que plantea el uso de algoritmos en decisiones judiciales. Argumenta que, aunque la IA puede ser una herramienta útil para asistir a los jueces en tareas como la búsqueda de precedentes legales o la predicción de resultados, nunca debe reemplazar la capacidad humana para interpretar y aplicar la ley. La autora subraya que los algoritmos utilizados en el ámbito judicial deben ser transparentes, explicables y auditables, para garantizar que no comprometan la imparcialidad ni los derechos de las partes involucradas. Además, destaca la importancia de establecer salvaguardas que aseguren que los jueces tengan la última palabra en las decisiones judiciales, evitando una dependencia excesiva de los sistemas automatizados.

Por su parte, Joan Guanyabens, explora en "Salut/IA" el impacto de la IA en el sector sanitario, centrándose en el caso de Cataluña. Guanyabens describe cómo la IA se está utilizando para mejorar la atención al paciente, optimizar recursos y avanzar en la investigación médica. Sin embargo, también identifica desafíos técnicos y éticos asociados al uso de datos médicos sensibles, como la privacidad, la seguridad y la equidad en el acceso a los beneficios de la IA. Entre las propuestas de Guanyabens destaca la creación de infraestructuras de datos que permitan un acceso seguro y eficiente a la información médica, así como el desarrollo de sistemas algorítmicos que sean inclusivos y representativos de la diversidad de pacientes. Además, subraya la necesidad de establecer mecanismos de supervisión que permitan identificar y corregir sesgos en los diagnósticos y tratamientos generados por algoritmos.

El caso práctico del Ayuntamiento de Barcelona, presentado por Paula Boet y Michael Donaldson en su aportación titulada "Datos, inteligencia artificial y servicios públicos: la apuesta del Ayuntamiento de Barcelona por la transparencia algorítmica y la protección de los

Deusto Journal of Human Rights
ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, 385-392
https://doi.org/10.18543/djhr.3201 • http://djhr.revistas.deusto.es/

derechos de la ciudadanía", demuestra cómo las administraciones locales pueden liderar iniciativas de transparencia y rendición de cuentas. Este capítulo detalla cómo el Ayuntamiento ha implementado registros públicos de algoritmos y ha desarrollado protocolos internos para regular su uso. Los autores destacan que estas iniciativas no solo fortalecen la confianza ciudadana, sino que también posicionan a Barcelona como un modelo de referencia en la gobernanza de la IA.

El volumen colectivo *Algoritmos abiertos y que no discriminen en el sector público* constituye un oportuna y equilibrada orientación acerca del *status quaestionis* del uso de los sistemas de IA en el sector público, combina rigor académico con propuestas prácticas, proporcionando un análisis integral de los retos a los que nos enfrentamos. Asimismo, incorpora el tratamiento de la más reciente evolución tanto normativa, como jurisprudencial y doctrinal, lo que nos permite afirmar, sin riesgo de incurrir en error, que se trata de una referencia inexcusable de quienes quieran iniciarse en el conocimiento de este ámbito jurídico, y que será bien recibida por legisladores, administradores y académicos interesados en construir políticas públicas responsables en la era digital, además de profesionales interesados en la interacción entre la tecnología y los derechos humanos. Con su enfoque claro y completo que centra su atención en la protección de los valores democráticos y la dignidad humana, esta obra no solo enriquece el debate teórico, sino que también ofrece herramientas prácticas para asegurar que la implementación de la IA en el sector público no comprometa los derechos fundamentales de los ciudadanos. No solo invita a reflexionar sobre los riesgos asociados a la IA, sino que también inspira a actuar para garantizar que estos sistemas sean diseñados e implementados con un profundo respeto por los derechos humanos y el bienestar social. Más allá de ser un análisis crítico, el libro es un llamado a la acción para construir un marco ético, legal y técnico que permita aprovechar las oportunidades de la IA sin comprometer los valores que sustentan nuestras sociedades democráticas.

José Miguel Iturmendi Rubia
CUNEF Universidad

Deusto Journal of Human Rights

Revista Deusto de Derechos Humanos

Deusto
Universidad de Deusto
University of Deusto