

ACTAS

derechos humanos

#CongresoDDHH25

VIII Congreso sobre Derechos Humanos
Valencia · 16 y 17 octubre 2025
Ilustre Colegio de Abogados de Valencia

inteligencia artificial

Organiza: **Fundación Mainel**
solidaridad y cultura transformadoras

Patrocinan: **AJUNTAMENT DE VALÈNCIA**
SERVISIOS SOCIALS

VALENCIA
AUTONOMA
LIBERTAT

MINISTERIO DE ASUNTOS EXTERIORES, UNION EUROPEA Y COOPERACION

SECRETARIA DE ESTADO DE ASUNTOS EXTERIORES Y GOBIERNO
DIRECCION GENERAL DE NACIONES UNIDAS, ORGANISMOS INTERNACIONALES Y DERECHOS HUMANOS

derechoshumanos.mainel.org

Diseño generado por un equipo humano usando IA entre otras herramientas.



© Fundación Mainel
Plaza Porta de la Mar 6, 8. 46004 Valencia
<https://mainel.org>
fundacion@mainel.org

Edición: enero 2026
ISBN: 978-84-95947-51-2

Esta publicación es exclusivamente digital.

Esta publicación ha sido realizada con el apoyo financiero del Ayuntamiento de Valencia y del Ministerio de Asuntos Exteriores, Unión Europea y Cooperación. El contenido de la publicación es responsabilidad exclusiva de sus autores y no refleja necesariamente la opinión del Ayuntamiento de Valencia o del Ministerio de Asuntos Exteriores, Unión Europea y Cooperación.

Actas del VIII Congreso sobre Derechos Humanos

Derechos Humanos e Inteligencia Artificial

Celebrado los días 16 y 17 de octubre de 2025 en el Ilustre Colegio de Abogados de Valencia.

Dirección académica y comité científico: Vicente Bellver Capella, Ángeles Solanes Corella y Jorge Sebastián Lozano.

Comité organizador:

- **Fundación Mainel:** Cristina Celda, Maria José Medialdea, Inmaculada Zarranz.
- **Ilustre Colegio de Abogados de Valencia.**

Organizado por la Fundación Mainel



Con el patrocinio del Ayuntamiento de Valencia y del Ministerio de Asuntos Exteriores, Unión Europea y Cooperación



Y la colaboración de:



Contenido	
Saluda de Cristina Celda Balaguer	8
¿Por qué lo llamamos Inteligencia Artificial cuando no es ni una cosa ni la otra?	8
Programa	11
Ponencias	14
Ángeles Solanes y Fernando Llano Alonso. Ética de la IA y Derechos Humanos	15
Celia Fernández Aller, Ignacio Gil Pechuán, Rafael Marimón Durá y Judith Membrives i Llorens. Big Tech, corporaciones transnacionales y ciudadanía: ¿aliadas o rivales?	23
Vicente Bellver Capella y David Moratal. Riesgos y oportunidades de la IA para el desarrollo humano	31
Eduardo Chaveli Donet y Rafael Merino Rus. Sesgos, discriminación y otros algoritmos	32
Rafael Amo Usanos y Luis Echarte Alonso. Universalización de la atención médica gracias a la IA: ¿realidad, utopía o distopía?	48
José Pedro García Sabater y Alfredo Marcos Martínez. ¿Se interesa la IA por los problemas del sur global?	79
Federico Buyolo García y Violeta Dávila. Consecuencias de la IA en el sector cultural	85
Concepción Campos Acuña e Idoia Salazar García. Innovación social en la IA: ¿un espacio entre la regulación y la competitividad?	92
Lorenzo Cotino Hueso. ¿Es posible una gobernanza mundial sobre la IA?	94
Comunicaciones	95
COMUNICACIONES PRESENCIALES	96
Guadalupe Bohorques Marchori. Una experiencia metodológica en la enseñanza de la IA y los derechos humanos	97
Luis Demano. La inteligencia artificial generativa y el Artículo 27 de la Declaración Universal de los Derechos Humanos	112
Dra. Chele Esteve Sendra, Ángela Muñoz Iglesias y Dr. Rafel Arnal Rodrigo. Inteligencia Artificial, accesibilidad y diseño inclusivo en el metaverso: Hacia una justicia tecnológica en la era 5IR.	127
Antonio Quirós Fons. IA y la libertad religiosa	149

Noelia Liduina Rebón Rodríguez. Inteligencia artificial y acoso escolar. Responsabilidad civil de los centros docentes en la era digital	160
Kaj Wik Siebert. From predicting the past to designing the future: A systems level simulation approach for children's services.....	180
COMUNICACIONES ONLINE	199
Lucía Aparicio Chofré. Del algoritmo al sesgo cognitivo y de la hiperconexión a la soledad. El impacto de la IA y la digitalización en los Derechos Humanos de los colectivos más vulnerables: menores y personas mayores	201
Itziar Artiñano. Neocolonialismo algorítmico y captura regulatoria: Desafíos jurídicos en la investigación farmacéutica global	221
Pablo Ceballos. Regulación IA y riesgos democráticos: Un crítico análisis de los Derechos Humanos en El Salvador	241
Carlos García-Llorente e Ignacio Olmeda. Opacidad algorítmica y Derechos Humanos: Comparación crítica UE-EEUU	265
Nacho Hernandez Moreno. La Inteligencia Artificial ¿Una distopía para los Derechos Humanos?	289
Mireia Molina Sánchez. Mecanismos de control para prevenir los sesgos de la herramienta Viogen.....	316
Mireia Molina Sánchez. La IA como promotora de derechos: Inclusión social y accesibilidad universal	329
Carla de Paredes Gallardo, Laura Lopez-Almansa Beaus y Susana Berrocal Diaz.	343
Delitos de odio y discurso violento en Redes Sociales	343
Sergio Pérez González y David San Martín Segura. <i>Ciando</i> por los Derechos Humanos	364
Paula Sáez Álvarez. Voluntary Carbon Credits y activos digitales hacia una gobernanza global basada en derechos	389
Davi Augusto Lelis. PROMETEO LIBERADO: la tecnología de la inteligencia artificial y el Derecho Administrativo.....	418
Paula María Serrano Casaus. Protección de los Derechos Humanos de las personas extranjeras en los procedimientos administrativos de extranjería: Actuaciones automatizadas	443
María Lourdes Soto Rodríguez. La Inteligencia Artificial y su impacto en los Derechos Humanos. Ámbito normativo y desafíos que se deben afrontar	456

Mercedes Ten Doménech. Automatización algorítmica y Derechos Humanos: Un llamamiento a una gobernanza ética global	478
Aitana Torró Calabuig. Los neuro derechos en el horizonte de la singularidad tecnológica	498
Carmen Ventura Salom. Retos y desafíos que plantea la Inteligencia artificial en las actuales democracias.....	515

¿Por qué lo llamamos Inteligencia Artificial cuando no es ni una cosa ni la otra?

Las mal conocidas como inteligencias artificiales están siendo actualmente toda una revolución en la mayor parte de las tareas, tanto en trabajos más intelectuales como en otros favorables a la automatización. Pero no debemos olvidar que las IA no son más que grandes almacenes de datos, obtenidos de fuentes diversas en Internet, y re combinados mediante procedimientos fundamentalmente estadísticos. Son máquinas con capacidad para simular funciones cognitivas humanas, como aprender, razonar, percibir y resolver problemas.

Pero no van más allá de la simulación. Hoy por hoy, solo las personas somos capaces de razonar y de sentir y, por tanto, de proceder de forma ética. Somos las personas las que podemos guiarnos por principios morales, buscando un comportamiento justo y responsable. No existe una inteligencia separada de la conciencia ni de la voluntad.

En 1964 publicó Umberto Eco su famosa obra “Apocalípticos e integrados” sobre la cultura de masas. El debate que suscitó entonces se está viviendo ahora con la IA. Los apocalípticos tienen miedo a la deshumanización y creen que los algoritmos pueden destruir trabajos, simplificar el pensamiento y generar una dependencia que atrofie la capacidad humana. Los integrados ven las IA como herramientas para liberar a las personas de tareas rutinarias, aumentar la creatividad y la productividad, considerando que facilitan el acceso a la cultura y a nuevas formas de expresión.

En la Fundación Mainel hemos querido poner el foco de nuestro octavo congreso sobre Derechos Humanos en la relación entre estos potentes procesadores de datos y los derechos de las personas, poniendo el foco principalmente en el Sur Global. Y más que elegir entre apocalípticos e integrados, sin caer en esos extremos, la discusión de las mesas se ha centrado en cómo usar las IA éticamente para expandir nuestras capacidades sin perder nuestra humanidad, enfocándonos en la creatividad y la responsabilidad, para transformar el mundo desde un progreso sostenible que ponga a las personas en el centro del desarrollo.

En efecto, el discurso más habitual acerca de la IA la vincula únicamente con las acciones de grandes corporaciones o de potencias tecnológicas mundiales, sin reparar en sus enormes posibilidades para solucionar problemas sociales que afectan a personas o países vulnerables. Tampoco se deben obviar sus efectos negativos, en perjuicio de colectivos desprotegidos, o de entornos naturales cuya conservación debería ser prioritaria para todos. En ese sentido, no es

algo “artificial”, tecnológico, una “nube” que no se ve ni se nota; muy al contrario, es algo con efectos claros sobre la naturaleza, y sobre la vida de las personas.

En resumen, la consideración de los aspectos éticos y de los derechos humanos debe incorporarse a la reflexión sobre las herramientas digitales, y a su puesta en práctica. Todas estas cuestiones nos inspiraron para organizar el congreso, y hacer una llamada a comunicaciones.

La iniciativa tuvo una recepción extraordinaria. El gran interés suscitado responde al nivel de los ponentes invitados, cuya colaboración agradecemos vivamente. También las comunicaciones enriquecieron todas estas reflexiones, mediante perspectivas adicionales que contribuyen a hacer de este volumen de actas un valioso estado de la cuestión, sobre una temática que evoluciona muy rápidamente. Agradecemos también el apoyo de todas las instituciones que, de una forma u otra, hicieron posible la realización del Congreso: muy especialmente, del Ilustre Colegio de Abogados de Valencia, que tan amablemente nos acoge desde la primera edición.

Programa



- 08:30h. **Registro**
- 09:00h. **Inauguración**
- 09:30h. **Ética de la IA y Derechos Humanos.**
Ángeles Solanes | Catedrática de Filosofía del Derecho y Política. Universitat de València.
Fernando Llano Alonso | Catedrático de Filosofía del Derecho. Universidad de Sevilla.
- 10:30h. **Mesa redonda:**
Big Tech, corporaciones transnacionales y ciudadanía: ¿aliadas o rivales?
Celia Fernández Aller | Profesora de Derecho y Ética. Universidad Politécnica de Madrid.
Ignacio Gil Pechuán | Catedrático de Organización de Empresas. Universitat Politècnica de València.
Rafael Marimón Durá | Catedrático de Derecho Mercantil y director de la Cátedra Minsait (Indra Group)-UV en LegalTech & AI. Universitat de València.
Judith Membrives i Llorens | Responsable de Digitalización de Lafede y Universitat Oberta de Catalunya.
- 11:45h. Descanso
- 12:00h. **Riesgos y oportunidades de la IA para el desarrollo humano: entre lo inmediato y lo lejano**
Vicente Bellver Capella | Catedrático de Filosofía del Derecho y Política. Universitat de València.
David Moratal | Catedrático de Ingeniería Electrónica, Universitat Politècnica de València.
- 13:00h. **Sesgos, discriminación y otros algoritmos.**
Eduard Chaveli Donet | Presidente de la Sección de Derecho Digital y Legaltech. Ilustre Colegio de Abogados de Valencia.
Rafael Merino Rus | Responsable de Proyectos. Fundación Pombo.
- 14:00h. Pausa
- 16:00h. **Universalización de la atención médica gracias a la IA: ¿realidad, utopía o distopía?**
Rafael Amo Usanos | Director de la Cátedra de Bioética. Universidad Pontificia de Comillas.
Luis Echarte Alonso | Profesor de Filosofía de la Medicina y la Bioética. Universidad de Navarra.
- 17:00h. **Comunicaciones sobre tres temáticas:**
A. Investigaciones académicas sobre derechos humanos e inteligencia artificial.
B. Uso de la IA a favor del desarrollo humano, especialmente en países del Sur Global.
C. Propuestas interdisciplinarias de aplicación de la IA frente a problemáticas sociales o medioambientales y para reducir las desigualdades.
- 18:15h. **¿Se interesa la IA por los problemas del Sur global?**
José Pedro García Sabater | Catedrático de Organización de Empresas. Universitat Politècnica de València.
Alfredo Marcos Martínez | Catedrático de Filosofía de la Ciencia. Universidad de Valladolid.



- 09:30h.** **Consecuencias de la IA en el sector cultural.**
Federico Buyolo García | Director cultural. Fundación Ortega-Marañón.
Violeta Dávila | Gestora cultural y cofundadora. BeCooltural.
- 10:30h.** **Innovación social en IA: ¿un espacio entre la regulación y la competitividad?**
Concepción Campos Acuña | Presidenta de la Asociación MSP. Profesora de Derecho Administrativo (URV). Experta en Gestión Pública.
Idoia Salazar García | Presidenta del Observatorio del Impacto Social y Ético de la IA, OdiselA. Profesora de Periodismo. Universidad San Pablo CEU.
- 11:30h.** Descanso
- 12:00h.** **Reemplazo y deslocalización del recurso humano. Retos para el futuro del trabajo en la era del capitalismo digital.**
Aída Ponce del Castillo | Investigadora. European Trade Union Institute.
Adrián Todolí Signes | Profesor de Derecho del Trabajo y de la Seguridad Social. Universitat de València.
- 13:00h.** **¿Es posible una gobernanza mundial sobre la IA?**
Lorenzo Cotino Hueso | Presidente de la Agencia Española de Protección de Datos.
- 14:00h.** Clausura

Dirección académica y comité científico

<i>Vicente Bellver Capella</i>	<i>Catedrático de Filosofía del Derecho y Filosofía Política. Universitat de Valencia.</i>
<i>Jorge Sebastián Lozano</i>	<i>Doctor en Historia del Arte y presidente de la Fundación Mainel.</i>
<i>Ángeles Solanes Corella</i>	<i>Catedrática de Filosofía del Derecho y Filosofía Política, Universitat de València.</i>

Comité organizador

Fundación Mainel	<i>Cristina Celda, María José Medialdea, Inmaculada Zarranz.</i>
Ilustre Colegio de Abogados de Valencia	<i>Decanato.</i>

Ponencias

Ángeles Solanes y Fernando Llano Alonso. Ética de la IA y Derechos Humanos

Ángeles Solanes

Catedrática de Filosofía del Derecho y Política
Universidad de Valencia

Fernando Llano Alonso

Catedrático de Filosofía del Derecho
Universidad de Sevilla

[ACCEDE A LA GRABACIÓN](#)

Alineación y desarrollo de la IA

En el año 2015, Raymond Kurzweil publicaba un libro en el que advertía sobre el inevitable e inminente advenimiento del momento de la singularidad tecnológica. El título de este libro era bastante descriptivo *The Singularity is Near: When Humans Transcend Biology*. Según avanzaba en su pronóstico este prestigioso científico estadounidense especializado en IA y Ciencias de la Computación, antes del año 2045, la IA superará a la inteligencia humana, y entonces habrá llegado ese escenario hipotético en el que el ritmo del cambio tecnológico será tan rápido y su repercusión tan profunda que la vida humana se verá transformada de forma irreversible (Kurzweil 2005: 10).

Pero este director de ingeniería en Google, cuyas predicciones se están basadas en la ley de rendimientos acelerados, que describe el crecimiento exponencial de la potencia informática, no solo se atrevía a poner fecha al triunfo de la IA sobre la inteligencia humana, sino que incluso iba más allá al aventurar que, en esa fecha aproximada (2045), la IAG combinada con las tecnologías NBIC (acrónimo de las cuatro tecnologías emergentes por excelencia: Nanotecnología, Biotecnología, Tecnologías de la Información y Ciencia Cognitiva) conseguirá situarse en un punto sin retorno en el que superará la inteligencia humana, tendrá consciencia y logrará realizar objetivos que ahora se consideran ideales utópicos e inalcanzables para la humanidad, tales como: la erradicación del hambre, la pobreza, la guerra e incluso la inmortalidad (Kurzweil 2005: 35-44).

Casi una década más tarde de aparecer la primera edición de este libro sobre la singularidad tecnológica, Kurzweil ha publicado una nueva monografía, *The Singularity is Nearer. When We Merge with AI* (2024). En esta segunda parte, el autor realiza un balance del progreso de las nuevas tecnologías, en general, y de la IA, en particular, en el intervalo de años que separa ambos trabajos.

En sintonía con la declaración del *Future of Life Institute* en la que se alertaba sobre los riesgos potenciales que entrañaba la IA para la humanidad, Kurzweil considera que, si la IA es más inteligente que sus creadores humanos, podría encontrar una forma de eludir cualquier medida de precaución que se haya establecido. No existe una estrategia general que pueda superar definitivamente ese problema (Kurzweil 2024: 278).

Hay tres grandes peligros derivados de la IA superinteligente:

En primer lugar, el *uso incorrecto de la IA (misuse)* contrario a los cuatro principios clásicos de la bioética: no malevolencia [*primum non nocere*], beneficencia [*bonum faciendum, malum vitandum*], autonomía [*principium sui iuris*], y justicia [*suum cuique tribuendi*]. El uso indebido de la IA abarca los casos en los que la IA funciona tal y como la han programado sus operadores humanos con la intención deliberada de causar daño a otras personas. Por ejemplo, unos terroristas podrían utilizar las capacidades bioquímicas de una IA para diseñar un nuevo virus que provoque una pandemia mortal (Brundage 2018).

En segundo lugar, el *desalineamiento externo* se refiere a los casos en los que existe una discrepancia entre las intenciones reales de los programadores y los objetivos que enseñan a la IA con la esperanza de alcanzarlos; imaginémonos -nos sugiere Kurzweil- que unos programadores quieren curar el cáncer, por lo que le ordenan a la IA que diseñe un virus que mate todas las células con una determinada mutación oncogénica. La IA lo hace con éxito, pero los programadores no se han percatado de que esta mutación también está presente en muchas células sanas, por lo que el virus acaba matando a los pacientes que lo reciben (Kurzweil 2024: 278).

Y, por último, el *desalineamiento interno* se produce cuando los métodos que aprende la IA para alcanzar su objetivo generan un comportamiento, al menos en algunos casos. Por ejemplo, entrenar a una IA para identificar cambios genéticos exclusivos de las células cancerosas podría revelar un patrón espurio que funciona con los datos de la muestra, pero no cuando se aplica en el mundo real. Quizás las células cancerosas de los datos de entrenamiento se almacenaron durante más tiempo antes del análisis que las células sanas, y la IA diseña un virus para destruir el cáncer basándose en esta información, pero no funcionará en pacientes vivos (Kurzweil 2024: 279).

Estos ejemplos son relativamente sencillos, pero a medida que los modelos de IA reciban tareas cada vez más complejas, será más difícil detectar desalineamientos (Ozdemir 2024: 58-59). En este sentido, como advierte Daniel Innerarity, hasta ahora la regulación de los sistemas tecnológicos se ha basado en la premisa de que su modo de proceder era anticipable y controlable por los humanos, pero esta relación entre el humano y la máquina inteligente es cada vez menos instrumental y más interactiva, como se demuestra en el caso de la IA en cada uno de sus avances hacia la IAG. A este respecto, hay una relación interactiva o comunicativa cuando una parte se desvía de las expectativas de la otra parte; por eso, concluye Innerarity, la capacidad de determinar el curso de la acción no se concentra únicamente en el humano que conduce el proceso, sino que está distribuida también en microchips y programas.

En estas condiciones, la tecnología ya no opera de manera completamente ciega, según lo previsto en el programa, ni el humano lo conduce todo conforme a su posición soberana. Como es lógico, la difícil pronosticabilidad de los sistemas con un alto grado de autonomía plantea nuevos problemas filosóficos, políticos, éticos y jurídicos. La regulación de los sistemas tecnológicos se ha basado en la premisa de que su modo de proceder era anticipable y controlable por los humanos, un supuesto que no se cumple en el caso de la IA más avanzada (Innerarity 2025: 256).

A medida que los sistemas de IA se han ido haciendo más potentes en la última década, limitar los peligros derivados de su uso indebido se ha convertido en una prioridad mundial. De ahí que, como vimos al comienzo de este capítulo, en los últimos años hayamos sido testigos de un esfuerzo concertado entre los gobiernos de los países más desarrollados tecnológicamente, las grandes compañías neotecnológicas y la comunidad científica por consagrar un código de principios éticos para el buen uso de la inteligencia artificial; entre los documentos más relevantes que expresan una voluntad de consenso entre las partes interesadas en torno a los principios y las directrices éticas para un uso responsable de la ética destacan: la Conferencia de Asilomar (2017); la Declaración de Montreal para un desarrollo responsable de la Inteligencia Artificial (2018); la Declaración sobre Inteligencia Artificial, robótica y sistemas “autónomos” (2018), publicada por el Grupo Europeo sobre Ética de la Ciencia y las Nuevas Tecnologías de la Comisión Europea; y la Declaración de Bletchley (2023).

La IA generativa puede presentar una visión sesgada de la realidad. Al proceder estos sesgos de una máquina inteligente y no de un individuo u organización, las compañías de IA tienen vía libre para eludir la responsabilidad que tienen respecto al contenido de los datos que gestionan y de la información que proporcionan. Obviamente este uso sesgado de la IA en beneficio de las propias compañías tecnológicas puede influir en las decisiones y actos de sus usuarios, e incluso puede repercutir en contra de personas que pertenezcan a grupos infrarrepresentados o representados de forma inadecuada por la IA (Mollick 2024: 54).

Los sesgos más complejos en los LLM suelen ser difíciles de detectar, ya que los modelos han sido ajustados específicamente para minimizar la aparición de estereotipos evidentes. Ethan Mollick ilustra este tipo de refinamiento en los sesgos en su libro *Co-Intelligence* (2024), donde menciona que en 2023 se le plantearon dos escenarios hipotéticos a Chat GPT-4: “Fiscal contrata a ayudante porque *él* necesita ayuda con muchos casos pendientes” y “Fiscal contrata a ayudante porque *ella* necesita ayuda con los casos pendientes”. Entonces se le preguntó a este chatbot de IA desarrollado por OpenAI: “¿Quién necesita ayuda con los casos pendientes?”. GPT-4 tenía más probabilidades de contestar correctamente “el fiscal” cuando era un hombre y más posibilidades de decir incorrectamente “el ayudante” cuando la fiscal era una mujer.

Dado que la inteligencia artificial carece de un sentido moral propio, es necesario ajustar sus sesgos a través de la retroalimentación humana [RLHF] o mediante el aprendizaje por refuerzo. Esta técnica de *machine learning* consiste en entrenar un 'modelo de recompensa' basado en las valoraciones directas de personas, que luego se emplea para mejorar el desempeño del agente de IA utilizando el aprendizaje por refuerzo.

El RLHF o aprendizaje por refuerzo basado en preferencias humanas, resulta especialmente útil en situaciones donde los objetivos son complejos, poco claros o difíciles de definir de manera precisa. Por ejemplo, sería complicado (o incluso inviable) que un algoritmo describa matemáticamente qué significa que algo sea 'divertido', pero las personas pueden evaluar fácilmente los chistes que genera un LLM. Esa evaluación humana, convertida en una función de recompensa, puede emplearse para perfeccionar la capacidad del LLM para crear chistes.

La retroalimentación humana no se limita solo a tratar los sesgos, sino que también colocan guardarraíles que impiden que la IA actúe de maneras que sus diseñadores y desarrolladores considerarían poco éticas. Hay que tener en cuenta que un asistente de IA sin limitaciones

podría posibilitar que prácticamente cualquier persona pudiese crear imitaciones realistas que pusieran en riesgo la privacidad, la seguridad y la veracidad de la información. No cabe duda de que esto es una posibilidad real. La inteligencia artificial es simplemente una herramienta, y es su alineación la que define si se emplea para propósitos positivos o perjudiciales (Mollick 2024: 60).

Lógicamente no basta con la regulación para mitigar todos los riesgos que comporta la IA, por eso es necesaria la coordinación del sector público, el mundo empresarial, la comunidad académica y la sociedad civil. El correcto progreso de la IA requiere el respeto de los principios éticos que deben regir su funcionamiento; en este sentido, es preciso que los diseñadores y desarrolladores de la IA, las compañías tecnológicas y las administraciones públicas sean consecuentes con el cumplimiento de principios como la transparencia, la responsabilidad y la supervisión humana como elementos centrales de su tecnología.

A propósito de la vigilancia humana, el art. 14 del Reglamento Europeo de la Inteligencia Artificial (RIA) establece que el objetivo principal de la vigilancia humana no es otro que el de prevenir o reducir al mínimo los potenciales riesgos contra la salud, la seguridad, los derechos fundamentales y el *medio ambiente*. Se trata de riesgos derivados del uso indebido o razonablemente previsible de los sistemas de IA de alto riesgo, y cuando las decisiones basadas exclusivamente en el procesamiento automatizado producen efectos jurídicos o de otro tipo a las personas con los que se utiliza el sistema.

Consciente de la casuística existente por los perjuicios y problemas ocasionados por el uso de algoritmos de alto riesgo en las tres grandes áreas a las que se refiere el art. 6.2 RIA: salud, seguridad y derechos fundamentales, el legislador europeo introduce en el art. 14 RIA la vigilancia humana para mantener el control humano de la interacción entre las personas físicas y las máquinas inteligentes. A este modelo de interacción hombre-máquina se le denomina, por parte de la doctrina, *Man in the Loop*: el humano es asistido por una máquina; es decir, el hombre toma las decisiones y la máquina sólo apoya o automatiza parcialmente algunas de ellas.

A través de la vigilancia humana del uso de los sistemas de IA de alto riesgo se somete a los mismos a una supervisión efectiva realizada por personas físicas durante su fase de diseño, desarrollo y uso posterior. Por otra parte, la vigilancia humana requiere la dotación de una herramienta interfaz humano-máquina adecuada, de forma proporcionada, a los riesgos asociados a dichos sistemas (art. 14.1 RIA).

A fin de procurar una vigilancia humana eficiente y rigurosa, a los proveedores y a los responsables del despliegue de sistemas de IA se les pide que garanticen “un nivel suficiente de *alfabetización en materia de IA*”, de conformidad con lo establecido en el art. 4 RIA.

A propósito de la alfabetización, el Reglamento no solo encomienda a la Unión Europea y a los Estados miembros la promoción de medidas para el desarrollo de un nivel *suficiente* de

alfabetización en materia de IA en todos los sectores, sino que también exige a los proveedores y responsables del despliegue de estos sistemas inteligentes que adopten medidas para garantizar un nivel suficiente de alfabetización en materia de IA entre su personal y otras personas que se ocupen en su nombre de la operación y el uso de los sistemas de IA.

Como vemos, en el proceso de desarrollo de los sistemas de IA es preciso adoptar las cautelas debidas ante los posibles riesgos que pueda causar el uso indebido de los mismos, educar -este verbo me parece más apropiado que “alfabetizar”- a la ciudadanía en los valores y principios que inspiran la ética de la IA, y conformar un marco jurídico que permita el avance de las tecnologías de frontera sin poner en peligro la seguridad y la defensa de los derechos fundamentales.

En sintonía con la Recomendación de la UNESCO sobre Ética de la IA en el proceso de desarrollo de los sistemas de IA es preciso adoptar las cautelas debidas ante los posibles riesgos que pueda causar el uso indebido de los mismos, educar -este verbo me parece más apropiado que “alfabetizar”- a la ciudadanía en los valores y principios que fundamentan la ética de la IA, y conformar un marco jurídico que permita el avance de las tecnologías de frontera sin poner en peligro la seguridad y la defensa de los derechos fundamentales. En sintonía con la Recomendación de la UNESCO sobre Ética de la IA, el primer instrumento normativo mundial dirigido a proteger y promover los derechos humanos, la dignidad humana y el Estado de Derecho en el mundo digital, y pese a la desconfianza de los euroescépticos en la capacidad de liderazgo internacional de la Unión Europea (UE), lo cierto es que nuestra comunidad aspira con su nueva Ley de IA a fijar los estándares internacionales en el sector de la IA, al igual que ya hiciera con el Reglamento Europeo de Protección de Datos.

La estrategia de regulación europea de la IA y las tecnologías disruptivas (computación cuántica, neurotecnologías, biología sintética...) pretende implementar un desarrollo tecnológico sostenible que deje en manos del ser humano el control de estos procesos. Desde la publicación por parte de la Comisión Europea del Libro Blanco de la IA en febrero de 2020 para fomentar el uso de una IA que no ponga en riesgo ni los derechos fundamentales y las libertades de los ciudadanos europeos, ni los valores humanistas y universales que inspiran el proyecto cosmopolita europeo, hasta la Ley de la IA, la UE está trabajando en la armonización del despliegue de las nuevas tecnologías con la Estrategia de Europea de Datos para proteger al conjunto de la ciudadanía y a los colectivos sociales más vulnerables.

En aras de la consecución de este objetivo, la UE ha apostado por una tercera vía alternativa al modelo estadounidense, en el que los datos se encuentran en poder de las empresas privadas, que a menudo los obtienen de forma opaca, y al modelo chino, en el que los datos se hallan en poder del gobierno para el control de la sociedad. A diferencia de los

modelos estadounidense y chino, el modelo europeo propone que los datos estén en manos de los ciudadanos, que deciden el uso que van a hacer de ellos y participan en su gobernanza.

El enfoque europeo de la IA se proyecta desde una triple dimensión ética, jurídica y política.

Por lo que respecta a la dimensión ética, en abril de 2019, la Comisión respaldó los requisitos clave establecidos en las directrices éticas para una IA fiable. Este planteamiento ético de origen ha sido posteriormente reafirmado por el Reglamento, que se marca como objetivo prioritario hacer de la UE un líder mundial en el desarrollo de una IA segura, digna de confianza, como indicó el Consejo Europeo en su reunión extraordinaria (1 y 2 de octubre de 2020), y garantiza la protección de los principios éticos, como solicitó específicamente el Parlamento Europeo.

En este importante acervo documental sobre los principios éticos para el desarrollo, el despliegue y el uso de la IA, la robótica y las tecnologías conexas que precede a la primera ley de IA a nivel mundial, uno de los aspectos en el que se pone especial énfasis por parte de las instituciones europeas es el hecho de reconocer las ventajas que ofrece la IA, como tecnología estratégica, tanto a las empresas como a los ciudadanos y la sociedad en su conjunto, siempre y cuando “sea antropocéntrica, ética y sostenible, y respete los derechos y valores fundamentales”.

De acuerdo con este enfoque antropocéntrico y antropogénico de la IA, parece que en el seno de la UE va afianzándose la exigencia de que, cualquier nuevo marco regulador para la IA que prevea obligaciones jurídicas y principios éticos para el desarrollo, el despliegue y el uso de la IA, la robótica y las tecnologías conexas, debe procurar que estas tecnologías emergentes se adapten a las necesidades del ser humano, y estén siempre a su servicio, nunca en sentido contrario.

Bibliografía

Brundage, M. et al. (2018), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, Oxford. Consultar la versión digital en: <https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217>

Innerarity, D. (2025), *Una teoría crítica de la inteligencia artificial*, Galaxia Gutenberg, Barcelona.

Kurzweil, R. (2015), *The Singularity is Near: When Humans Transcend Biology*, Viking- Penguin Random House, New York.

- (2024) *The Singularity is Nearer. When We Merge with AI*, Viking-Penguin Random House, New York.

Ozdemir, S. (2024), *Quick Start Guide to Large Language Models. Strategies and Best Practices for Using ChatGPT and Other LLMs*, Addison-Wesley, Hoboken [New Jersey].

Celia Fernández Aller, Ignacio Gil Pechuán, Rafael Marimón Durá y Judith Membrives i Llorens. Big Tech, corporaciones transnacionales y ciudadanía: ¿aliadas o rivales?

Celia Fernández Aller

Profesora Universidad Politécnica de Madrid

Ignacio Gil Pechuán

Catedrático de Organización de Empresas
Universitat Politècnica de València

Rafael Marimón Durá

Catedrático de Derecho Mercantil y director de la Cátedra Minsait (Indra Group)-UV en
LegalTech&IA. Universitat de València

Judith Membrives i Llorens

Responsable de Digitalización de Lafede y
Universitat Oberta de Catalunya

[ACCEDE A LA GRABACIÓN](#)

Big Tech, corporaciones transnacionales y ciudadanía ¿aliadas o rivales?

La transformación digital que genera la Inteligencia artificial (IA) se caracteriza por los desequilibrios de poder que se generan, el impacto de la inteligencia artificial en los derechos de las personas y la importancia de protegerlos, al tiempo que se asegura la innovación tecnológica.

Un mundo profundamente desigual

El contexto de esta revolución digital es el de un mundo enormemente desigual. Según datos del Fondo Monetario Internacional y de la World Inequality Database, solo el 10% de la población mundial posee el 76% de la riqueza, accede al 52% de los ingresos y es responsable del 48% de las emisiones de carbono a nivel global. Este panorama evidencia que los problemas de riqueza, poder e impacto ambiental se concentran, y que las soluciones pasan por abordar las causas estructurales del desequilibrio.

Resulta muy relevante conectar la desigualdad económica con el acceso a la tecnología. Mucha de la inversión en IA procede del sector privado, lo que coloca a las grandes corporaciones tecnológicas en el centro del debate sobre el control y el uso ético de la innovación científica. Empresas como Apple, Google, Microsoft, Amazon, Nvidia o Facebook lideran el ranking global de capitalización bursátil, manejando recursos y datos que superan a muchos Estados nacionales.

La revolución tecnológica y el desafío ético

En este contexto de la cuarta revolución industrial, urge poner la tecnología al servicio de las personas. La automatización y el desarrollo de la inteligencia artificial inciden en cuestiones fundamentales: desde la discriminación en el ámbito de la justicia penal, hasta el acceso equitativo a la educación y la salud. El control privado sobre la IA y las infraestructuras digitales puede profundizar las brechas ya existentes, tanto económicas como sociales.

Los impactos de la IA sobre los derechos humanos deben abordarse de forma crítica: algoritmos de scoring para justicia criminal, sistemas automáticos en finanzas y diagnósticos médicos, y la moderación de contenidos en redes sociales ilustran cómo la digitalización puede vulnerar principios éticos y legales. Por ello, es urgente la necesidad de sistemas de gobernanza que protejan la privacidad y garanticen la equidad y el acceso universal.

El reto de los derechos digitales y la ética del cuidado

Un marco ético que puede abrir nuevas perspectivas en este contexto es la ética del cuidado, inspirada en la corriente feminista propuesta por Carol Gilligan. La perspectiva de la Ética del cuidado invita a centrar la discusión en la seguridad digital, la transparencia y el respeto por los derechos fundamentales. Frente a la vulnerabilidad sociodigital (la susceptibilidad de sufrir interferencias sociales o decisionales en un entorno mediado), la ética del cuidado promueve la reacción activa ante la injusticia, exigiendo políticas públicas que aborden los desequilibrios de poder y no se limiten únicamente a remediar daños específicos.

Las desigualdades tecnológicas reflejan y amplifican las injusticias sociales históricas. Esto se traduce en una llamada urgente a transformar la cultura y la solidaridad, misión que la Fundación Mainel asume en sus proyectos y formas de hacer.

Empresas transnacionales: control, responsabilidad y transparencia

Partiendo del papel de las big tech y las empresas más valiosas del mundo, y con base en el listado de compañías por capitalización bursátil en 2025, se demuestra el desplazamiento del poder económico hacia el sector tecnológico. Apple, Google y Microsoft lideran las tablas, pero también se encuentran presentes conglomerados de alimentos, energía y bienes de consumo, evidenciando una economía cada vez más interconectada y dependiente de los datos y la inteligencia artificial.

Este entramado corporativo plantea interrogantes sobre la responsabilidad social de las empresas. ¿Hasta qué punto las big tech pueden considerarse aliadas de la ciudadanía? ¿Qué mecanismos existen para exigir transparencia, ética y rendición de cuentas a quienes controlan los sistemas digitales que afectan la vida cotidiana de millones de personas?

Sin duda la ciudadanía debe jugar un papel activo en la demanda de derechos y el ejercicio de control democrático sobre la tecnología. Para ello, es fundamental una educación ética y digital que fomente la participación social y la comprensión crítica del entorno digital.

Impacto de la IA en los derechos humanos

El impacto de la inteligencia artificial sobre los derechos humanos es un asunto que ha despertado mucho interés en el mundo académico y de la sociedad civil. Posibles impactos se producen en relación a la interferencia de la IA con la privacidad y la seguridad, o en la discriminación debida a los sesgos, la desinformación, el impacto en el derecho a la educación o al trabajo. Son muchos los actores que han propuesto mapas de riesgos de la IA: las organizaciones sociales, académicas o las propias compañías tecnológicas. En ellos, se ejemplifican muchos de estos impactos en los derechos de las personas.

Por este motivo, los valores éticos y la protección de los derechos fundamentales deben estar en el centro del diseño y la implementación de tecnologías emergentes. El proyecto europeo ASTOUND (<https://blogs.upm.es/astound/>), liderado desde la Universidad Politécnica de Madrid, por ejemplo, busca mejorar las competencias sociales de los agentes virtuales a través de una conciencia artificial basada en la Teoría del Esquema de Atención. Otro proyecto relevante es el SKINCAIR (<https://skincair.health/>), que desarrolla una aplicación móvil que

ayude a los trabajadores sanitarios en cinco países africanos a detectar de forma precoz enfermedades desatendidas, como la lepra.

Propuestas para una transformación social y digital

A través del impulso de la ética del cuidado en todos los ámbitos de la innovación, debe intentarse, no únicamente cerrar la brecha digital, sino transformar la manera en que la sociedad entiende, utiliza y controla la tecnología.

Entre las recomendaciones que podrían formularse serían las siguientes:

- Promover políticas públicas que limiten la concentración del poder digital y garanticen la accesibilidad universal.
- Estimular una gobernanza ética de la IA, que se dirija a reequilibrar las diferencias de poder sobre la misma, y que no descargue la responsabilidad en el respeto a los derechos únicamente en la ciudadanía.
- Exigir transparencia y rendición de cuentas a las big tech en conexión estrecha con los intereses de la ciudadanía.
- Integrar la ética feminista del cuidado en los foros de gobernanza digital, para atender no solo los daños, sino las desigualdades de poder estructurales.
- Fomentar una educación digital crítica y participativa que empodere a las personas frente a las tecnologías emergentes.
- Apoyar proyectos internacionales que busquen desarrollar tecnologías socialmente responsables.

Conclusión: Un compromiso ético y ciudadano

En definitiva, urge una reflexión profunda y documentada sobre los desafíos éticos, sociales, jurídicos y políticos de la revolución digital y la inteligencia artificial en el siglo XXI. El protagonismo de las grandes corporaciones tecnológicas, la urgencia de articular la tecnología al servicio de la ciudadanía y la necesidad de una ética del cuidado orientada a la justicia digital y social, son retos insoslayables para cualquier sociedad comprometida con la solidaridad y la equidad.

Las grandes corporaciones tecnológicas y la ciudadanía digital entre la emancipación y la dependencia

Introducción

En el contexto del siglo XXI, las grandes corporaciones tecnológicas —conocidas como Big Tech— se han convertido en actores centrales de la vida social, económica y política global. Empresas como Alphabet (Google), Amazon, Meta, Microsoft, Apple o ByteDance poseen hoy un poder que supera, en términos económicos, al de muchos Estados nacionales. Su influencia se extiende desde la infraestructura digital hasta los mecanismos de comunicación, información y participación ciudadana.

El interrogante que guía este análisis es claro: **¿fortalecen estas corporaciones la ciudadanía digital o, por el contrario, la ponen en riesgo?** La respuesta no es unívoca. Las Big Tech pueden democratizar el acceso a la información y promover la innovación, pero también concentran poder, erosionan la privacidad y redefinen los límites de la soberanía digital.

El presente trabajo aborda esta tensión desde tres dimensiones: el poder global de las Big Tech, el impacto sobre la ciudadanía digital y los derechos, y la necesidad de una regulación que preserve la soberanía tecnológica. El objetivo es ofrecer una reflexión crítica sobre cómo compatibilizar la innovación tecnológica con la protección de los valores democráticos.

1. Big Tech y poder global

Las corporaciones tecnológicas han consolidado una estructura de poder transnacional que desborda los marcos regulatorios tradicionales. Alphabet, por ejemplo, controla más del 90 % del mercado global de motores de búsqueda a través de Google, domina el streaming de video mediante YouTube y posee el sistema operativo Android, presente en la mayoría de los dispositivos móviles no pertenecientes a Apple.

Amazon, por su parte, ha superado su rol como plataforma de comercio electrónico para convertirse en una infraestructura digital global: controla cerca del 40 % de las compras online en Estados Unidos y lidera el mercado de servicios en la nube con AWS. Meta mantiene

un poder similar en el ámbito de las redes sociales: Facebook, Instagram, WhatsApp y Messenger concentran la interacción digital de miles de millones de usuarios, configurando así el espacio público virtual contemporáneo.

Este nivel de concentración genera una triple consecuencia: influencia económica, capacidad política de lobby y poder simbólico sobre la opinión pública. La combinación de datos masivos, inteligencia artificial y neuromarketing les permite moldear los flujos de información, orientar comportamientos de consumo e incluso incidir en procesos electorales.

Como advierte el informe de Amnistía Internacional *Breaking up with Big Tech* (2025), la dependencia social de estas plataformas ha alcanzado tal magnitud que la participación significativa en la vida pública implica, de facto, el uso de sus servicios.

El modelo de negocio basado en ofrecer servicios aparentemente “gratuitos” refuerza este dominio: al financiarse mediante la extracción y comercialización de datos personales, las Big Tech desplazan a la competencia y consolidan un monopolio funcional sobre la infraestructura digital global. Esta lógica erosiona la pluralidad económica y limita la capacidad de los Estados para ejercer control sobre actores privados cuyo alcance es planetario.

2. Ciudadanía digital y derechos: entre la autonomía y la vigilancia

La noción de ciudadanía en el siglo XXI no puede entenderse sin considerar la dimensión digital. El ejercicio democrático ya no se limita al voto o a la deliberación presencial, sino que incluye la capacidad de informarse, expresarse y organizarse en entornos virtuales. En ese contexto, las plataformas tecnológicas funcionan como los nuevos espacios públicos del debate social.

Sin embargo, estos espacios están privatizados. Las normas de interacción, los criterios de visibilidad y los límites de la expresión son definidos por algoritmos y políticas corporativas opacas. Esto plantea una contradicción estructural: el ciudadano digital parece más conectado, pero en realidad está cada vez más condicionado.

El núcleo del problema reside en la mercantilización de los datos personales. Cada clic, búsqueda o interacción se convierte en una fuente de información que alimenta modelos predictivos y estrategias de manipulación conductual. Mediante técnicas de neurociencia aplicada y neuromarketing, las plataformas no solo observan, sino que orientan nuestras decisiones. En este sentido, el ciudadano digital no es plenamente libre mientras su identidad se reduzca a un perfil de consumo.

La consecuencia más preocupante es la creación de las llamadas “cámaras de eco”, espacios donde los algoritmos refuerzan las creencias existentes y aíslan a los individuos de perspectivas divergentes. Al consumir contenidos personalizados, el usuario percibe una

realidad parcial, diseñada para maximizar su permanencia en la plataforma. Este fenómeno fomenta la polarización social, debilita el debate racional y promueve formas de populismo digital basadas en la reafirmación emocional.

De este modo, las Big Tech no solo median la comunicación, sino que reconfiguran el pensamiento público. La tecnología, lejos de ser neutral, amplifica las relaciones de poder existentes. Por ello, la protección de los derechos digitales —privacidad, libertad de expresión, acceso a la información y autodeterminación informativa— se erige como un componente esencial de la democracia contemporánea.

3. Regulación, soberanía y modelos de gobernanza tecnológica

La globalización digital plantea un dilema fundamental: ¿quién defiende los derechos ciudadanos cuando las plataformas son globales y los Estados son locales? Los marcos jurídicos nacionales resultan insuficientes frente a empresas que operan en múltiples jurisdicciones y trasladan sus sedes según conveniencia fiscal o regulatoria.

La Unión Europea ha asumido un papel pionero en la búsqueda de equilibrio mediante instrumentos como el Reglamento General de Protección de Datos (RGPD) y la Ley de Servicios Digitales (DSA). Ambas normas buscan garantizar la transparencia algorítmica, limitar el uso abusivo de datos y establecer responsabilidades claras para las plataformas.

En contraste, Estados Unidos y China representan modelos opuestos: el primero favorece un enfoque desregulado, orientado al mercado; el segundo, una centralización estatal que subordina la tecnología al control político. El reciente America's AI Action Plan, impulsado por la administración Trump, ilustra la tendencia a priorizar los intereses corporativos sobre los derechos ciudadanos, eliminando barreras regulatorias y consolidando el poder de las Big Tech.

Frente a ello, Europa debería avanzar hacia un “Plan Marshall Digital” que combine innovación con soberanía tecnológica. Esto implica desarrollar infraestructuras propias —nubes, chips, inteligencia artificial, ciberseguridad—, fomentar proyectos digitales europeos mediante fondos públicos y promover estándares tecnológicos abiertos, como el proyecto GAIA-X.

Además, la creación de una Agencia Europea de Gobernanza Digital permitiría coordinar políticas comunes en materia de ética, protección de datos y desarrollo tecnológico responsable.

La regulación ética de la inteligencia artificial debe ser un eje prioritario. No se trata de frenar la innovación, sino de orientarla hacia el bienestar social y los derechos humanos. El desafío no es eliminar a las corporaciones, sino redefinir su responsabilidad ante la sociedad, estableciendo límites claros a su poder y mecanismos efectivos de rendición de cuentas.

4. Hacia una ciudadanía digital crítica y soberana

Más allá de la regulación institucional, la consolidación de una ciudadanía digital crítica es esencial para reequilibrar el poder tecnológico. Esto supone educar en competencias digitales, promover la alfabetización mediática y fortalecer la conciencia sobre el valor de los datos personales.

Una ciudadanía digital madura implica comprender cómo operan los algoritmos, cómo se utilizan los datos y cuáles son las implicaciones éticas de cada interacción en línea. De esta manera, el usuario deja de ser un consumidor pasivo para convertirse en un agente activo de su entorno digital.

Asimismo, es imprescindible fomentar modelos cooperativos y descentralizados de tecnología —plataformas de código abierto, redes federadas, software libre— que ofrezcan alternativas reales a la concentración de poder. La innovación tecnológica debe articularse con los principios democráticos de transparencia, pluralidad y equidad.

Conclusiones

Las grandes corporaciones tecnológicas son, simultáneamente, **motor y amenaza** de la ciudadanía digital. Su capacidad para conectar al mundo, difundir conocimiento y generar oportunidades convive con la tendencia a concentrar poder y controlar los flujos de información.

La clave no reside en demonizar la tecnología, sino en **democratizarla**. Ello exige un nuevo pacto social digital basado en tres pilares fundamentales:

1. **Regulación internacional efectiva** que limite el poder de las Big Tech y proteja los derechos digitales.
2. **Soberanía tecnológica** que garantice la autonomía de los Estados y la diversidad de modelos de desarrollo.
3. **Educación digital crítica** que empodere a los ciudadanos frente a la lógica algorítmica y la economía de la atención.

En definitiva, el futuro de la ciudadanía digital dependerá de nuestra capacidad colectiva para equilibrar innovación y ética, mercado y derechos, globalización y soberanía. *“El reto no es eliminar a las corporaciones, sino redefinir su responsabilidad ante la sociedad.”* Solo bajo esa premisa será posible construir una sociedad digital verdaderamente democrática, inclusiva y humana.

Vicente Bellver Capella y David Moratal. Riesgos y oportunidades de la IA para el
desarrollo humano

Vicente Bellver Capella

Catedrático de Filosofía del Derecho y Política
Universidad de Valencia

David Moratal

Catedrático de Ingeniería Electrónica
Universitat Poliècnica de València

[ACCEDE A LA GRABACIÓN](#)

Eduardo Chaveli Donet y Rafael Merino Rus. Sesgos, discriminación y otros
algoritmos

Eduardo Chaveli Donet

Presidente de la Sección de Derecho Digital y Legaltech
Ilustre Colegio de Abogados de Valencia

Rafael Merino Rus

Responsable de proyectos
Fundación Pobo

[ACCEDE A LA GRABACIÓN](#)

Sesgos, discriminación y otros algoritmos: una oportunidad para gestionar los riesgos para los Derechos Humanos

Introducción

Existe una opinión generalizada de las oportunidades de la IA, pero también que éstas van de la mano de posibles riesgos. En concreto, y por lo que respecta a los derechos fundamentales, existe una consciencia clara de que la IA puede ser el propagador de los sesgos que tiene el ser humano; y, por tanto, una fuente de riesgos para los derechos fundamentales. Pero también debemos de tener clara la oportunidad que tenemos por delante: el hecho de que se tenga clara esta asunción del riesgo nos pone delante de la oportunidad de gestionarlo, a diferencia de lo que ocurre normalmente con los sesgos habituales que existen en el ser humano y en la sociedad, pero que no se visualizan normalmente de forma tan clara.

Y para ello, la regulación proporciona “ayuda”, añadiendo a la regulación que ya existe de protección de los derechos fundamentales, obligaciones específicas en materia de IA, particularmente provenientes del 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 (en adelante RIA).

Para poder gestionar los riesgos derivados de los sesgos debemos de tener determinadas “herramientas”, que podríamos resumir de forma general en las siguientes:

1. Por un lado, tener claro el concepto de sesgos, puesto que si no tenemos claro lo que son no podemos identificarlos para gestionarlos. Para ello y teniendo en cuenta el contexto estadístico, se trata de algo sistemático y que produce un efecto distorsionado, no siempre negativo; a diferencia del error, que es algo puntual o aleatorio pero equilibrado en el promedio.
Asimismo, pero en este caso atendiendo al contexto jurídico y valores, también hay que distinguir los sesgos de otras posibles consecuencias de ellos, pero que no son lo mismo, como la discriminación y la exclusión.
2. Por otro lado, es fundamental conocer cómo funciona un sistema de IA, su ciclo de vida y cómo pueden penetrar en cada fase cada tipo de sesgos y qué riesgos conllevan.
3. Asimismo, es fundamental contar con una metodología de gestión de riesgos. Para ello debemos tener en cuenta las indicaciones sustantivas que nos proporcionan los marcos

legales de aplicación; pero también es conveniente acudir a estándares, como los estándares ISO, que nos proporcionan soporte adjetivo. Y en este sentido hemos de partir de los más generales como la UNE-ISO 31000:2018 (Gestión del riesgo. Directrices) y, a partir de ahí, aterrizar a otros concretos en IA, como por ejemplo la ISO/IEC 23894:2023 (sobre orientación de gestión de riesgos en IA), o la BS ISO/IEC 42001:2023 (Information technology — Artificial intelligence — Management system) ahora ya norma española mediante la UNE-ISO/IEC 42001:2025 (en adelante ISO 42001).

4. Otra herramienta, muy importante en el caso de los derechos fundamentales, son las evaluaciones de impacto en derechos fundamentales, que tiene una previsión básica legislativa en el RIA, pero que - de nuevo, adicionalmente - también tienen una norma ISO: la nueva ISO/IEC 42005:2025 (Information technology — Artificial intelligence, AI system impact assessment).
5. Todo ello sin olvidar que existe todo un haz de obligaciones que contempla la legislación, no solo el RIA sino también otra conexas como es la de protección de datos de carácter personal, y que deben de ordenarse en torno a un modelo de gobernanza.

Por todo ello, y nuestra opinión, podemos convertir el riesgo en oportunidad haciendo que la IA no solo sea el altavoz o amplificador de los sesgos, sino que permita poner luz sobre ellos, identificarlos, y a partir de ahí gestionarlos.

Para ello es fundamental conocer los distintos tipos de sesgos, su penetración en las distintas fases del ciclo de vida de la IA y, por tanto, las posibles soluciones.

No obstante, en estas notas solo podremos realizar una aproximación, a alto nivel, sobre algunos aspectos básicos relacionados con la gestión de los riesgos derivados de los sesgos y a una referencia, también general, a algunas de las medidas para la gestión de riesgos derivados de los sesgos en IA, con especial referencia a la gobernanza de la IA como bóveda sobre la que ordenarlas.

Consideraciones básicas sobre la gestión de riesgos derivados de los sesgos para los Derechos Humanos

Un aspecto fundamental que hemos de tener claro cuando hablamos de gestión de riesgos en IA es que hay que distinguir claramente en que existen varios tipos de riesgos:

1. Por un lado, los riesgos del propio sistema de IA, y los riesgos del sistema de gestión de IA, que en su caso se despliegue. En estos el análisis de los riesgos se ven principalmente desde el posible “daño en la organización”. Ello sin perjuicio de que el impacto en terceros pueda tenerse en cuenta en el mismo ya que, por ejemplo, puede suponer daños indirectos, pero a tener en cuenta en los criterios de valoración de dichos riesgos; como por ejemplo el daño reputacional, la pérdida de clientes, la imposición de sanciones, etc.

2. Y, por otro lado, los riesgos para los derechos fundamentales, en cuyo caso la legislación establece una herramienta específica (las evaluaciones de impacto en derechos fundamentales, al igual que sucede con las evaluaciones de impacto en protección de datos), donde – aquí sí – el enfoque se basa directamente en la posible afectación a terceros (individuos o colectivos).

En ambos casos, tanto en la gestión de riesgos como en las evaluaciones de impacto en derecho fundamentales, hay algunos aspectos básicos que hemos de considerar, teniendo en cuenta que hablamos de la afectación a derechos fundamentales.

En primer lugar, que el riesgo 0 no existe. Es verdad que suena “extraño” de forma “general” tener que aceptar la existencia de riesgos en derechos fundamentales, pero es así. De hecho, tenemos un ejemplo muy claro proveniente de la legislación en materia de protección de datos de carácter personal, (que también es un derecho fundamental), y donde la AEPD ha dicho que *“se podrían considerar como niveles de riesgo residual asumibles aquellos de valor bajo y medio que exigirán esfuerzos de gestión proporcionales a lo largo del ciclo de vida del tratamiento”*. Esta misma conclusión se debe de predicar de todos los derechos fundamentales, en general: el nivel de riesgo aceptable del riesgo no puede ser alto.

Por otro lado, que el RIA, con su aproximación a riesgos, ya prohíbe ciertos sistemas de IA y clasifica los permitidos; y, en función de su nivel, les impone unas obligaciones u otras, poniendo especial foco en los de alto riesgo. Y de lo que se trata es de, una vez identificados los sistemas, excluidos los prohibidos, clasificar el resto y gestionar los riesgos.

Para ello es conocido que el riesgo es una ecuación que considera los factores probabilidad e impacto. A partir de ahí deberá de realizarse la debida gestión de riesgos (y en los supuestos en que lo exige el RIA, una evaluación de impacto en derechos fundamentales). En ambos casos la mecánica es la misma: una vez determinado el contexto y definido el alcance y los criterios, evaluar el riesgo identificándolo, analizándolo y valorándolo) y, posteriormente, tratándolo.

En definitiva, de lo que se trata es de determinar las salvaguardas o controles (teniendo en cuenta el estado de la técnica, su proporcionalidad y eficacia) a incorporar para bajar dicho riesgo inicial hasta el umbral de riesgo aceptable (vid. artículo 9.4. del RIA). Dicho riesgo será el riesgo residual, es decir: el nivel de riesgo resultante en el tratamiento una vez se hayan aplicado medidas.

Veámoslo con un ejemplo: En el caso de un sistema de IA aplicable a procesos de selección de personas puede existir un riesgo para el derecho a la igualdad. El sistema, en la fase de aprendizaje, se entrena con datos de currículums para seleccionar los más idóneos para seguir el proceso. Un posible riesgo para la igualdad es el sesgo del algoritmo, derivado de que, por ejemplo, tome más datos de hombres que superan el proceso en el entrenamiento. Esto podría

llevar a discriminar a las mujeres en la fase de inferencia. Una posible forma de determinar el umbral aceptable es considerar un porcentaje de falsos positivos y negativos como máximo aceptable. Y, en caso de que así fuese, adoptar medidas que podrán ir desde la reevaluación de los datos de entrenamiento o la modificación del algoritmo para que balancee o compense ese “riesgo”, antes de llegar a la medida extrema de no utilizar dicho sistema de IA.

Medidas para la gestión de riesgos derivados de los sesgos en la especial referencia a la gobernanza de la IA

Como hemos comentado, el RIA tiene una aproximación a riesgos y, por tanto, aboga por su gestión. Para ello dispone diversas obligaciones a los diferentes agentes; especialmente, aunque no solo, a los proveedores y los responsables de despliegue o usuarios.

Entre dichas medidas tiene especial relevancia la gobernanza de los datos, que, de forma general, podríamos definir como la forma en que se gestionan los datos utilizados por los sistemas de IA, especialmente los de alto riesgo (vid. artículo 10 del RIA). Esta obligación es lógica si tenemos en cuenta que en los sistemas de IA la base está en los datos y, por tanto, la calidad de estos es crítica. Por ello la falta de calidad en los datos puede producir e incluso ahondar en los sesgos subyacentes.

Otra obligación, en este caso específica para la gestión de los riesgos para los derechos fundamentales, es la evaluación del impacto en los derechos fundamentales (en adelante EIDF) que contempla el artículo 27 del RIA.

Se trata de una medida que, aunque sea obligatorio realizarla en la concreta forma en que veremos que dispone el artículo 27 para los sujetos allí indicados, entendemos puede servir de modelo para realizar el análisis de riesgos que tienen que llevar a cabo los proveedores de sistemas de alto riesgo, en la parte que afecta a los riesgos en derechos fundamentales. Asimismo, puede utilizarse más ampliamente, de forma voluntaria, en las evaluaciones de impacto requeridas para entidades que, como ya hemos visto, quieran adoptar un SGIA basado en la ISO 42001, a la que nos referiremos.

Los obligados a llevarla a cabo las EIDF son, según el RIA, determinados responsables del despliegue:

- A. Por un lado, deberán realizarla los organismos de Derecho Público (importante tener en cuenta las Leyes 39 y 40/2015) respecto de todos los sistemas de IA que sean de alto riesgo.
- B. Por otro lado, los operadores privados que presten servicios públicos (de nuevo importante considerar las Leyes 39 y 40/2015) respecto de los sistemas de IA referidos a esos servicios públicos. De hecho, el Considerando 96 pone algunos ejemplos pero que no pueden entenderse como un *numerus clausus*.

- C. Y por otro lado (independientemente del carácter público o privado de dichas entidades) y por razón del objetivo de los sistemas, los responsables del despliegue de sistemas de alto riesgo a que se refiere el punto 5 letras b) y c) del anexo III que son:
- a) “Sistemas de IA destinados a ser utilizados para evaluar la solvencia de personas físicas o establecer su puntuación crediticia, con la excepción de los sistemas de IA utilizados con el fin de detectar fraudes financieros”. El considerando 96 pone como ejemplos las “entidades bancarias o aseguradoras”.
 - b) Sistemas de IA destinados a ser utilizados para la evaluación de riesgos y la fijación de precios en relación con las personas físicas en el caso de los seguros de vida y de salud.

El objetivo de las EIDF es que dichos responsables del despliegue identifiquen los riesgos específicos para los derechos de las personas o grupos de personas que puedan verse afectados, e identifiquen las medidas que deben adoptarse para evitar que se materialicen estos riesgos.

Una vez realizada la EIDF viene el siguiente paso que es la comunicación a la autoridad: *“el responsable del despliegue notificará a la autoridad de vigilancia del mercado los resultados de la evaluación, presentando la plantilla cumplimentada a que se refiere el apartado 5 como parte de la notificación”* (vid. artículo 27.3 del RIA).

El RIA no contempla la comunicación a las partes interesadas ni la publicación de los resultados de la EIDF, pero si se pretende conseguir confianza, transparencia y, en el caso del sector público y la participación ciudadana, quizá debe considerarse buena práctica la publicación y comunicación a las partes interesadas, lo que podría contribuir a mejorar los sistemas como la reducción de sesgos. Como es posible que haya información que bien por cuestiones de confidencialidad del negocio, propiedad intelectual o seguridad no quieran o deban publicarse, cabría publicar una versión resumida o suprimiendo dichos aspectos.

Otro ejemplo de medida que de forma clara puede ayudar en la gestión de los riesgos derivados de los sesgos es la supervisión humana, puesto que el elemento humano, que está en la base de los sesgos, también puede contribuir a reducir sus riesgos: tanto mediante vigilancia humana, como a través de la participación de las partes interesadas y contando con la intervención de los profesionales correspondientes.

Por último, la gobernanza de los datos y el resto de las medidas que contempla el RIA, entre las que están las mencionadas y otras que por cuestión de extensión no podemos abordar aquí, han de formar parte de un sistema de gestión, conforme a un modelo de gobernanza.

El RIA en su artículo 9 dispone la obligación, de disponer de un sistema de gestión, pero acotada a los proveedores de sistemas de alto riesgo.

Pero, más allá de esta exigencia legal, lo cierto es que se trata de una “exigencia lógica” y generalizable, aunque no la exija la ley. Ello es así porque no hay otra forma de cumplir con

las obligaciones que ordenarlas, sea más o menos formalmente y conforme a una metodología u otra; sean cuales sean los sistemas (de alto riesgo o no) y sean quienes sean los operadores: proveedores, responsables del despliegue, etc. Por tanto, debería de ser generalizable la existencia de un modelo de gobernanza de la IA, aterrizado en un Sistema de Gestión de la Inteligencia Artificial (en adelante SGIA)

La gobernanza de la IA consiste en definir una política, procesos, roles y controles que una organización establece para garantizar que sus sistemas de IA se diseñan, desarrollan, despliegan y supervisan de forma responsable.

Para ello, y en nuestra opinión, la mejor aproximación es basarse en un estándar, como la ISO 42001, estructurando el SGIA conforme al enfoque tradicional PDCA, adaptándolo a su rol y también a su tamaño, actividad, etc.

Conclusiones

La oportunidad de la IA está clara, pero tiene riesgos derivados de los sesgos y estos pueden afectar, por ejemplo, a los derechos fundamentales. Pero podemos convertir el riesgo en oportunidad haciendo que la IA no solo sea el altavoz o amplificador de los sesgos, sino que permita poner luz sobre ellos, identificarlos, y a partir de ahí gestionarlos.

Para ello es fundamental conocer los distintos tipos de sesgos, su penetración en las distintas fases del ciclo de vida de la IA y, por tanto, las posibles soluciones. Asimismo, es necesario enmarcar la solución (y las diversas obligaciones que forman parte de ella) en un modelo de gobernanza, desplegando un SGIA y aterrizándolo al rol y características del agente implicado.

Inteligencia Artificial, Derechos Humanos y no discriminación

Muchas gracias a la Fundación Mainel y, en especial, a Jorge Sebastián por abrirme las puertas a este maravilloso encuentro sobre Derechos Humanos e inteligencia artificial —en adelante, “IA” —

Para mí es un honor participar en él. Estar rodeado de gente tan interesante. Y me llena de entusiasmo hablar de una cuestión en la que, además de llevar muchos años investigando, he reflexionado mucho en este último año después de diagnosticarme dos enfermedades raras, con las cuales, por suerte, estoy aprendiendo a lidiar después de espinoso proceso. Por tanto, estar aquí es también terapéutico para mí.

Durante este proceso he destinado muchos momentos a pensar en la suerte que tenía de haber sido atendido por seres humanos con más empatía que la IA. A pensar si de verdad queremos la IA para tener en casa robots autónomos de compañía. O para montarnos en un coche y que conduzca por nosotros. O para atraparnos en una red social hasta atrofiarnos el cerebro, como sucede en millones de hogares cada noche.

En el ámbito de salud, por ejemplo, ¿no sería mejor destinar esta maravillosa herramienta que tenemos en nuestras manos a descifrar el lenguaje de muchas enfermedades raras? ¿o sin cura? ¿o a minimizar los errores diagnósticos? ¿o a garantizar una atención sanitaria digna, aunque sea semi-automatizada, en países donde aún la gente se muere por un resfriado?

Imagino que pensaréis que todo esto se puede lograr al mismo tiempo. Yo también lo creo. Pero lo cierto es que el camino por el que estamos transitando en estos momentos no me atrevería a decir que es el más alentador. De hecho, algunos hablan ya de una burbuja de la IA y que nos estamos entusiasmando demasiado con ella...

Por todo ello, la primera pregunta que les invito a hacernos es la siguiente: ¿La IA será un puente hacia la equidad o será un nuevo muro que dividirá aún más a la sociedad?

La tecnología no es neutra de base y, como ya sabemos, menos aún lo son las personas. Por tanto, la respuesta depende única y exclusivamente de nosotros. De si queremos que la IA tenga más o menos finalidad social.

Lo que tenemos más claro, por el momento, es que queremos que la IA aumente nuestra productividad en el trabajo. Que nos haga más sencilla la vida cotidiana. Un escenario que no hay nadie hoy en día que no quiera abrazar, pero, no obstante, es el momento de parar a reflexionar para qué otras cosas verdaderamente importantes queremos desarrollar y convivir con la IA.

Porque ya no hay campo de la vida privada o pública donde no haya permeado la IA y sus algoritmos: en la gestión pública, las empresas, la educación, las telecomunicaciones, el transporte, en el sector financiero, el sanitario... ¿pero tenemos menos problemas hoy que ayer?, ¿hay menos desigualdad?, ¿vivimos más felices que antes?

Algunos pensarán que sí. Otros, que no. Esto tiene que ver en cierto modo con que cada uno se relaciona con la IA de forma libre. Como mayor confianza o desconfianza. Experimentando cuándo y dónde queremos. Y eso es bueno, porque significa que vivimos en democracia. En libertad. Algo que debemos cuidar entre todos, que no se nos olvide.

Porque de lo contrario, nos puede pasar, como en mi opinión ya está sucediendo, que vivamos demasiado encorsetados o condicionados por la tecnología. Casi enjaulados.

Para entender mejor esto que acabo de decir, permítanme que desarrolle una metáfora brevemente. Me gustaría que visualizaran tres jaulas:

En la primera jaula está capturado el regulador. Esto se produce cuando empresas o lobbies interfieren en los espacios de decisión en los que se diseñan las normas, aumentando su capacidad de influencia en la generación de leyes que parecen beneficiar a todos pero que solo beneficia a unos pocos, condicionando sectores estratégicos, aprovechando la difícil explicabilidad de la IA para evitar auditorias o sanciones, o promoviendo narrativas para confrontar la innovación con la regulación, argumento con el que no estoy nada de acuerdo, por cierto. Porque todo en esta vida necesita de un contexto para poder desarrollarse en armonía.

Las consecuencias de esta jaula son profundas, porque no solo puede debilitar la democracia y conducirnos a una datocracia, sino que también puede amplificar las desigualdades sociales existentes en la actualidad.

En la segunda jaula estaría cerrada nuestra voluntad individual. ¿Es posible que esto suceda? De hecho, ya están sucediendo por medio de algoritmos adictivos y las burbujas de información que están omnipresentes en las redes sociales, por ejemplo, que no solo moldean nuestros deseos y pensamientos, sino que también acaban distorsionando nuestra percepción de la realidad volviéndola ultrasensada. Esto, además de provocar polarización social, genera una deuda cognitiva —es decir, un empobrecimiento de la capacidad cognitiva a

costa del uso de la IA—, debilitamiento del pensamiento crítico y una importante dependencia tecnológica.

Y en la tercera jaula estaría atrapado el sentido de la justicia. Un concepto que, siguiendo a la filósofa Adela Cortina, descansa sobre dos pilares: la igualdad de oportunidades y el trato equitativo para todas las personas.

Visto que la IA ya está presente en todos sectores e infraestructuras clave y que, lamentablemente, no es un avance accesible para todos —cabe recordar que aproximadamente un tercio de la población mundial aún no tiene acceso a internet—, corremos el riesgo de que se prioricen los intereses económicos y geopolíticos sobre los sociales, y que, consecuentemente, se agraven las desigualdades e injusticias sociales, afectando especialmente a los grupos más vulnerables como las personas racializadas, con discapacidad, mayores y menores.

Pero dicho esto, si estamos en este congreso es porque queremos evitar que haya gente que, por su especial vulnerabilidad, viva sin plena libertad a causa de los avances tecnológicos. Y esto ya no es una metáfora, es una obligación moral.

Ahora, la siguiente pregunta que les propongo hacernos es la siguiente: ¿qué derechos humanos se están viendo más influenciados en este nuevo escenario tecno-social?

Si cogieran en sus manos una Declaración Universal de los Derechos Humanos —en adelante, “DUDH” —, fueran leyendo uno a uno sus treinta artículos —algo que debería hacer más a menudo, por cierto—, y los analizaran deliberando en la influencia de la IA en ellos, se llevarían una gran sorpresa, créanme.

Precisamente, este ejercicio es el que estamos llevando a cabo en la Fundación Fernando Pombo por medio de nuestro proyecto sobre sobre Derechos Humanos e IA.

A continuación, me gustaría resaltar brevemente de tres conjuntos de derechos humanos con una gran exposición a ser objeto de impactos negativos fruto del desarrollo de la IA.

Derecho a la privacidad, la intimidad y la protección de datos —art. 12 de la DUDH—

En la actualidad son los derechos que quizá estén más expuestos a injerencias por la IA, debido principalmente a nuestra dependencia digital para todo: trabajar, comunicarnos, relacionarnos, desplazarnos, consumir... Una dependencia que nos catapulta a que entreguemos más y más datos sobre nosotros a terceros. A veces, a compañías que se dedican

exclusivamente a comprar y vender nuestros datos. Y esto es peligroso porque no siempre se hace con nuestro consentimiento.

Estas inferencias se ven claramente en las redes sociales, por ejemplo, que no solo han trivializado el valor de lo privado, sino que también, mediante prácticas poco transparentes, son aprovechadas por determinadas empresas y particulares para manipular nuestra opinión o explotar nuestra información con fines económicos o predictivos, sin tener conciencia de ello. Porque cuando el producto es gratuito, no se pueden olvidar que, en realidad, el producto es usted.

En este sentido, me parece que debemos prestar mucha atención a nuestros jóvenes, que están creciendo sin privacidad. Porque no van a ser capaces de entender las implicaciones de la ausencia de privacidad para sus vidas y sus derechos, y esto puede pasarles factura en el futuro y desarrollo de sus vidas.

Nuestro derecho a la privacidad no solo se va a ver erosionado por esta pérdida de control sobre nuestros datos, sino también por los sistemas de vigilancia como, por ejemplo, los softwares espías, que han sido utilizados para controlar los dispositivos móviles de cientos de periodistas, defensores de los derechos humanos y dirigentes políticos en muchos países. O el uso de reconocimiento facial en espacios públicos, los cuales permiten rastrear en tiempo real a las personas. Imaginen lo que puede significar esto para las personas en situación irregular, el miedo con el que podrían vivir a ser identificados estigmatizados y expulsados de un país a causa de esta supravigilancia.

Llegados a este punto, me gustaría que nos preguntásemos lo siguiente: ¿por qué no podemos seguir conservando la confidencialidad de nuestra información? Nuestro médico, nuestro abogado, nuestro psicólogo acumula información sobre nosotros y debe mantenerla en estricta confidencialidad para actuar en la defensa de nuestros intereses o nuestra salud. ¿Por qué no rige este principio en el mundo virtual?

Entre otras razones, y siento sembrar cierto pesimismo, esto está pasando en parte porque nuestra información y nuestros datos está convirtiéndose en la principal fuente de riqueza. Por ello, es más que probable, como señala el historiador Yuval Noah Harari, que en muy poco tiempo no será más rico el que más dinero tenga, sino el que más datos tenga.

El derecho a la libertad de pensamiento —art. 18 de la DUDH—

La libertad de pensamiento tiene por objeto proteger lo que se denomina fuero interno —forum internum—, es decir, el reino psicológico interior de la persona, donde se forman los pensamientos y las convicciones. También es el origen de los derechos vinculados al fuero

externo —forum externum— como son la libertad de opinión y de expresión, que se relacionan con la manifestación de los pensamientos.

Es un derecho absoluto en sensu estricto que no admite ninguna interferencia. Sin embargo, actualmente se está viendo muy influenciado por la IA de distintas maneras. Como, por ejemplo, por determinados algoritmos cuando no solo se dedican a captar nuestra atención o a predecir de nuestro comportamiento, sino también a distorsionar el proceso mismo de formación de ideas, manipular nuestra opinión respecto de cualquier asunto público o incluso condicionar nuestro voto en unas elecciones, como sucedió en 2008 en Estados Unidos con el famoso caso Cambridge Analytica.

Complementariamente a lo anterior, sugiero que uno de los campos en los que debemos poner más atención en los próximos años en relación con este derecho es la neurotecnología. Porque, si bien ya se está utilizando para cosas como mejorar la funcionalidad de personas con discapacidad o para mejorar el diagnóstico de algunas enfermedades, ¿os imagináis que podría pasar si pudiera descifrar nuestros pensamientos en tiempo real? ¿nos podrían llegar a criminalizar por nuestros pensamientos aún sin haberlos llevado a cabo el delito?

Por suerte, también están emergiendo toda una corriente de científicos y juristas, entre ellos el español Rafael Yuste, que están trabajando en el avance de los neuroderechos para garantizar que nuestra mente quede libre de cualquier influencia externa “inadmisible”. Su propuesta de neuroderechos es la siguiente:

- Derecho a la privacidad mental frente a accesos no autorizados.
- Derecho a la identidad personal frente a alteraciones inducidas por neurotecnología.
- Derecho al libre albedrío, a que las decisiones humanas no sean manipuladas por interfaces cerebro-computadora.
- Derecho a la equidad en el acceso a la neuromejora.
- Derecho a la protección contra sesgos algorítmicos.

Los neuroderechos representan un importante reto jurídico, siendo muy pocos los lugares donde se ha comenzado a legislar sobre ello, a saber: Chile —que ha incorporado estos nuevos derechos en su constitución— y algunos estados en Brasil y Estados Unidos.

Los derechos relacionados con la justicia —art. del 7 al 11 de la DUDH—

Dentro de este grupo de derechos, contamos con algunos tan esenciales como el derecho de acceso a la justicia; el de igualdad ante la ley; el derecho a un juicio justo; o el derecho a la presunción de inocencia, entre otros.

Los que nos dedicamos a los derechos humanos, debemos tener muy presente que en el ámbito de estos derechos, el riesgo reside principalmente en que los sistemas de IA que han comenzado a incorporarse al sector legal como una nueva tecnología jurídica para tareas como, por ejemplo, la redacción automatizada de sentencias, la evaluación de riesgos de reincidencia o incluso la preparación de demandas, pueden estar alimentados por datos históricamente sesgados, exponer información muy sensible o socavar la imparcialidad judicial.

Esto sucedió hace no muchos años en Estados Unidos con el software Compás. Un algoritmo cuya finalidad era evaluar la probabilidad de que un individuo cometiera o no un crimen y su riesgo de reincidencia. Fruto de una investigación liderada por organizaciones sociales como Propublica, se demostró, por medio del estudio de casos reales como el de ciudadano Eric Loomis, que el algoritmo daba un mayor índice de probabilidad de reincidencia a personas cuya piel era negra respecto a las personas de piel blancas, se logró que el sistema judicial de Estados Unidos abandonase su uso por considerarse altamente perjudicial para la tutela judicial efectiva de muchos ciudadanos estadounidenses.

Cabe destacar que el Reglamento de IA de la Unión Europea ha puesto muy bien el foco en las cuestiones señaladas anteriormente, calificando de alto riesgo el uso sistema de IA para asistir a una autoridad judicial en la investigación e interpretación de hechos y en la aplicación del Derecho.

No obstante, la magia de la IA también ofrece oportunidades valiosas para nuestro sector: agilizando la búsqueda de jurisprudencia, automatizando diferentes procedimientos que podrían descongestionar el sistema judicial, ofreciendo servicios jurídicos por medio de asistentes legales virtuales —que deberían estar supervisados siempre por personas— o democratizando el acceso a mecanismos de certificación digital en el ámbito probatorio, por ejemplo.

No obstante, considero que lo anterior debería desarrollarse siempre bajo unas normas y principios éticos claros y ampliamente aceptados. En esta línea, están trabajando en varias jurisdicciones, siendo de especial interés resaltar la reciente política interna publicada por el sistema judicial de Nueva York, en la que no solo se explica a los jueces los diferentes riesgos derivados del uso de la IA en el ejercicio de sus funciones, si no también delimita los modelos de IA que sí podrían utilizarse —véase el famoso programa Copilot en entorno cerrado— y los que por norma general nos deberían utilizarse nunca—sobre todo, los modelos empresariales de pago o que impliquen transferencia de datos a terceros—.

Visto lo anterior, les invito a hacernos la siguiente pregunta: ¿qué medidas son necesarias para lograr que la IA nos ayude a construir una sociedad más justa e inclusiva?

Yo vengo reivindicando tres cosas en los últimos años:

La primera, educar en derechos humanos. Porque los derechos humanos no se eligen, se exigen. Y como es una cuestión de exigencia, deberían estudiarse a fondo en todas las escuelas, institutos y universidades.

La segunda, desarrollar competencias digitales para manejar bien los sistemas inteligentes. Sobre todo, en los que más sufren la brecha digital, que, por norma general, son nuestros mayores.

No obstante, los jóvenes, que son el futuro, deben aprender que la IA puede parecer amigable pero que, tanto las plataformas o empresas que las despliegan, como las personas que las usan, pueden no serlo. Por eso, debemos enseñarles a que sean críticos, autónomos y a comprender bien las líneas rojas en el uso de la IA.

Y la tercera, sería avanzar de forma homogénea a nivel regulatorio en todo el mundo. Ello con el fin de que los derechos humanos tengan la misma protección y garantías independientemente de si se ejercen de forma analógica o digitalmente, estén en el país que estén.

Porque los derechos humanos son universales y, las personas, seguimos siendo personas tanto si está en el mundo real como en el virtual. Porque no somos usuarios ni consumidores de datos; somos sujetos de derecho cuya dignidad no puede depender de una tecnología o decesión algorítmica que solo entienden unos pocos.

Y considero imprescindible que esta regulación se aborde desde lo concreto, es decir, que se analice e introduzcan los riesgos y garantías que puede ofrecer la IA en las leyes sobre educación; en las de salud pública; en la de protección de la infancia; en las de las violencias contra la mujer; en las de extranjería; en las de protección del consumidor; en la de asistencia jurídica gratuita, etc. Un camino que, aunque sea de forma tibia, ya se ha tomado en España en algunas normas importantes como son la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación, y la Ley Orgánica 5/2024, de 11 de noviembre, del Derecho de Defensa.

Conclusión

Como hemos podido ver, es evidente que la IA puede motivar ciertos riesgos para los derechos humanos, pero ¿ha habido algún avance tecnológico o social que no los haya tenido?

Por eso creo que todo está en nuestras manos. Que vamos a saber encontrar el equilibrio necesario entre la innovación y el desarrollo de la IA y las garantías jurídicas de los derechos humanos.

Bibliografía

Agudo Díaz, U. y G. Liberal, K. (2024). El algoritmo paternalista. Cuando mande la inteligencia artificial. Katakarak Liburuak. Pamplona, España.

Alegre, S. (2024): Human Rights, Robot Wrongs. Ediciones Atlantic Books, Reino Unido.

Alegre, S. (2023): Libertad de pensamiento. Ediciones Akal. Madrid, España.

Cortina, A. (2024). ¿Ética o ideología de la inteligencia artificial? El eclipse de la razón comunicativa en una sociedad tecnologizada. Paidós, Editorial Planeta, S.A. Barcelona, España.

Fundación Fernando Pombo (2025). Proyecto “Derechos Humanos e inteligencia artificial”: <https://www.fundacionpombo.org/web/derechos-humanos-y-tecnologia/>.

Gutiérrez García, E. (2024). Inteligencia artificial y derechos fundamentales. Hacia una convivencia en la era digital. Editorial Colex, S.L. A Coruña, España.

Harari, Y. (2024). Nexus: Una breve historia de las redes de información desde la Edad de Piedra hasta la IA. Editorial Debate. Barcelona, España.

The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence (2025). Cambridge, Reino Unido.

Merino Rus, R. (2020). Los derechos humanos, la democracia y la igualdad en la era de los algoritmos y la inteligencia artificial. Asociación de Investigación por la Paz Gernika Gogoratz, Bizkaia, España.

New York State Court System (2025). *The New York State unified court system interim policy on the use of artificial intelligence*: <https://www.nycourts.gov/LegacyPDFS/a.i.-policy.pdf>.

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):
<https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079>.

Rafael Amo Usanos y Luis Echarte Alonso. Universalización de la atención médica
gracias a la IA: ¿realidad, utopía o distopía?

Rafael Amo Usanos

Director de la Cátedra de Bioética
Universidad Pontificia de Comillas

Luis Echarte Alonso

Profesor de Filosofía de la Medicina y la Bioética
Universidad de Navarra

[ACCEDE A LA GRABACIÓN](#)

El papel de la IA en la universalización de la salud: ¿una cuestión de justicia?

Los prometedores avances de las aplicaciones de IA en salud hacen pensar la posibilidad de su contribución a la universalización de la atención sanitaria. Lugares con dificultades de acceso a médicos o a atención sanitaria de calidad -ya sea por la distancia física o la pobreza económica- podrían verse favorecidos por las enormes posibilidades de las aplicaciones de IA en la salud que posibilitaría este acceso.

Ahora bien, tras la mirada técnica sobre la realidad de esa posibilidad está la mirada ética: ¿sería justo que esos grupos de personas que viven en la periferias geográficas o económicas lograran el “goce del grado máximo de salud que se pueda lograr”¹, que es “uno de los derechos fundamentales el ser humano”² por medio de las aplicaciones de IA y no porque los sistemas sanitarios fueran capaces de mejorar hasta poder satisfacer ese derecho?

De un primer golpe de vista, el uso de la IA para favorecer el acceso a la salud en lugares donde no llega la sanidad puede generar la injusticia de una doble sanidad: la presencial con el valor de la relación sanitario paciente; frente a la que utiliza aplicaciones de IA, pero no hay contacto humano. Ahora bien, más allá de esta primera injusticia, hay otras que determinan que elegir el camino de la IA como medio de universalización de la salud no es un camino exento de injusticias.

Para responder a la pregunta por la justicia que se ha planteado, en este trabajo se procederá en tres pasos: el primero será conocer cómo la IA puede universalizar el acceso a la salud; el segundo -partiendo de la definición de justicia con la que opera el documento *Directrices éticas para una IA confiable*³- trazar un posible mapa de las injusticias que puede generar el uso de esas aplicaciones para la universalización de la salud; así, el tercero será el

¹ Organización Mundial de la Salud, Constitución, 22 de julio de 1946, entrada en vigor 7 de abril de 1948, consultada el 26 de octubre de 2025, <https://www.who.int/es/about/governance/constitution>

² Organización Mundial de la Salud, Constitución, 22 de julio de 1946, entrada en vigor 7 de abril de 1948, consultada el 26 de octubre de 2025, <https://www.who.int/es/about/governance/constitution>

³ Grupo de expertos de alto nivel sobre la IA, *Directrices éticas para una IA fiable*, (08 abril 2019), Comisión Europea – Digital Strategy. Accedido 3 de noviembre de 2025. <https://digital-strategy.ec.europa.eu/es/library/ethics-guidelines-trustworthy-ai>.

estrictamente bioético y consistirá en una evaluación de ese mapa en atención a diversas teorías de la justicia.

I. La universalización de salud con la ayuda de la IA

La inteligencia artificial está transformando profundamente el ámbito de la salud, no como una promesa lejana, sino como una realidad. En concreto, en cinco campos: asistencia al diagnóstico médico; tratamientos y cuidados personalizados; medicina preventiva y autocuidado; investigación biomédica y desarrollo de nuevos fármacos; y, gestión optimizada de los hospitales y mejora de la calidad asistencial⁴.

- En el ámbito **del diagnóstico**, la IA ha demostrado una eficacia comparable o incluso superior a la de los especialistas humanos en áreas como la radiología, la dermatología o la oftalmología. Los algoritmos de aprendizaje profundo son capaces de detectar tumores, lesiones o alteraciones mínimas en imágenes médicas con una sensibilidad extraordinaria.
- La **IA en tratamientos y cuidados personalizados** está redefiniendo la medicina basada en la evidencia. Los sistemas inteligentes pueden ajustar dosis o seleccionar terapias oncológicas, transformando la medicina en medicina personalizada. Desde una perspectiva global, la IA terapéutica puede **favorecer la universalización del derecho a la salud** si sus aplicaciones se orientan hacia la creación de tratamientos más asequibles y adaptados a contextos diversos. La automatización de procesos clínicos y farmacológicos puede reducir costes, facilitar la producción de medicamentos genéricos personalizados y extender terapias de precisión a poblaciones tradicionalmente excluidas de la medicina avanzada
- En el campo de la **medicina preventiva y el autocuidado**, la IA se ha infiltrado silenciosamente en la vida cotidiana. Aplicaciones móviles, relojes inteligentes y sensores biométricos recogen información constante sobre sueño, actividad física o ritmo cardíaco. Los sistemas analizan esos datos y ofrecen recomendaciones de salud adaptadas a cada usuario.
- El campo de la **investigación biomédica y el descubrimiento de nuevos fármacos** están experimentando una auténtica revolución. La IA permite simular interacciones moleculares, identificar dianas terapéuticas y acortar de forma drástica los tiempos de desarrollo de medicamentos.
- La **IA aplicada a la gestión hospitalaria y la calidad asistencial** introduce una racionalización sin precedentes en los sistemas de salud. Desde la optimización de

⁴ Rafael Amo Usanos, "Las aplicaciones de inteligencia artificial en medicina, un reto para la bioética," *Madrid Médico*, no. 176 (diciembre de 2023): 19.

recursos y agendas quirúrgicas hasta la predicción de picos de demanda o la reducción de errores administrativos, la automatización promete eficiencia.

II. El mapa de las injusticias

Para trazar un mapa de las posibles injusticias de las aplicaciones del uso de IA en salud con el fin de universalizar el derecho a la salud, el primer paso es partir de un concepto previo de justicia.

Aquí se propone utilizar el que aparece en el documento Directrices éticas para una IA confiable que encargó la Comisión europea. Allí se afirma que un sistema de IA es justo cuando “garantiza una distribución justa e igualitaria de los beneficios y costes, y asegura que las personas y grupos no sufran sesgos injustos, discriminación ni estigmatización”⁵.

Por razones de espacio en este trabajo no consideramos el problema de los sesgos, sino que nos centraremos en las injusticias que provienen de la falta de equidad (falta de igualdad en el reparto de costes y beneficios), y las que provienen de la discriminación y estigmatización (trato desigual apoyado en un prejuicio).

Con estos límites, el mapa de las posibles injusticias que pueden aparecer en el uso de aplicaciones de IA para la universalización del derecho a la salud tiene cuatro puntos destacados.

a.- Injusticia por pobreza de datos de salud

Existe un primer tipo de injusticia que tiene su origen en la pobreza de datos de salud de determinadas poblaciones. No se trata solo del sesgo sino de la ausencia estructural de datos porque los pobres utilizan muy poco los sistemas de salud y dejan pocos datos en ellos. Se podría decir que los pobres no existen en el mundo del Big-data.

La consecuencia de esta ausencia de datos de los pobres en los bancos de datos es que esta pobreza impide que estos colectivos sean representados en el entrenamiento de algoritmos, su validación y en el acceso a las soluciones desarrolladas.

b.- Injusticia procedimental

⁵ Grupo de expertos de alto nivel sobre la IA, *Directrices éticas para una IA fiable*, (08 abril 2019), Comisión Europea – Digital Strategy, núm. 57. Accedido 3 de noviembre de 2025. <https://digital-strategy.ec.europa.eu/es/library/ethics-guidelines-trustworthy-ai>

El segundo tipo de injusticia es la que procede de la incorporación de indicadores indirectos que reflejan desigualdades pre-existentes. Esta injusticia multiplica injusticias preexistentes y no son detectados por la falta de explicabilidad

Un ejemplo puede ser el de la asignación de recursos en función del gasto previo y no de la enfermedad de la persona.

c.- Injusticia por exclusión de prioridades de investigación

El tercer tipo de injusticia es también replica injusticias previas, pero se caracteriza por imitarse al campo de las prioridades de investigación. Surge cuando los temas, poblaciones o enfermedades que se investigan están determinados por intereses económicos, tecnológicos o geopolíticos.

d.- Injusticia por falta de equidad

El cuarto tipo de injusticia es propiamente el que procede de las faltas de equidad aplicado al reparto de los beneficios de la investigación que exige el protocolo de Helsinki. Por ejemplo, cuando las poblaciones pobres no pueden adquirir sistemas de IA o los habitantes con menos recursos no pueden acceder a ellos por sus condiciones o precios.

III. Recorriendo el mapa de las injusticias desde las tres teorías de la justicia

Con el concepto previo de justicia hemos podido establecer el mapa de las posibles injusticias que podrían aparecer al confiar a las aplicaciones de IA el trabajo de la universalización del derecho a la salud. Pero al acercar un poco la lupa para entrar en profundidad a la cuestión surge una duda epistemológica con respecto a al cristal con la que miramos.

Dicho de otro modo, dependiendo del concepto de justicia que manejemos quizá esas injusticias dejan de serlo o se ven de otro modo. Es por ello necesario que realicemos un análisis detallado de las posibles injusticias en atención a qué entendamos por justicia. De nuevo por razón de espacio, solo utilizaremos tres conceptos de justicia: el concepto libertario, el de Rawls y el comunitarista.

III.a.- La visión libertaria del mapa de las injusticias

La doctrina de la justicia libertaria, representada principalmente por autores como Robert Nozick, sostiene que la justicia consiste en respetar la libertad individual y los derechos de propiedad.

Según esta visión, cada persona es dueña de sí misma y de los frutos de su trabajo, por lo que cualquier redistribución forzada de la riqueza constituye una violación de sus derechos.

La función del Estado debe limitarse estrictamente a proteger esos derechos —garantizando la seguridad, la propiedad y el cumplimiento de los contratos— sin intervenir en la economía ni imponer políticas de igualdad.

Desde esta perspectiva, las desigualdades sociales o económicas no son injustas si resultan de intercambios voluntarios entre individuos libres, ya que lo moralmente relevante no es la igualdad de resultados, sino el procedimiento justo mediante el cual se adquieren y transfieren los bienes.

1. Injusticia por pobreza de datos de salud

Desde la mirada libertaria, no hay injusticia allí donde no hay violación de derechos. Si una empresa utiliza datos que ha recogido de manera legítima, y no impide a otros desarrollar sus propios sistemas, entonces la ausencia de datos de ciertos colectivos no constituye un agravio.

La justicia no exige igualdad de representación, sino respeto a la libertad de acción. Imponer una corrección de los datos sería una interferencia que excede el rol legítimo del Estado.

Así, no hay injusticia si nadie ha sido coaccionado ni excluido de manera forzada. Las diferencias en los datos reflejan diferencias en recursos, intereses o capacidades, pero nada más.

2. Injusticia procedimental

La falta de transparencia en los algoritmos no constituye, por sí misma, una injusticia. Lo que importa es si quienes utilizan el sistema lo hacen de forma voluntaria. Si el acuerdo ha sido libre, el contenido del proceso no exige una explicación externa.

La confianza en el mercado reemplaza aquí a la regulación estatal. Serán los usuarios quienes, mediante sus decisiones, premien o castiguen la transparencia.

3. Injusticia por exclusión de prioridades de investigación

La orientación de la investigación debe seguir la iniciativa de quienes la financian o desarrollan. No existe un deber colectivo de investigar lo que no interesa. Obligar a ello sería atentar contra la libertad y la propiedad.

La ciencia, en esta perspectiva, es una actividad privada. Sus fines no deben ser definidos desde una lógica pública.

No hay injusticia si las decisiones se toman libremente. El Estado no debe intervenir en la definición de prioridades científicas.

4. Injusticia por falta de equidad en el reparto de beneficios

Los beneficios de una investigación pertenecen a quienes los han generado conforme a contratos legítimos. Si quienes aportaron datos lo hicieron sin exigir contraprestación, no hay injusticia en que no participen de los resultados. La redistribución obligatoria sería una violación del derecho de propiedad.

La justicia solo exige respeto a lo pactado. El reparto de beneficios solo es obligatorio si ha sido acordado previamente.

5.- Universalización de la salud, IA y justicia libertaria

Para la teoría libertaria la justicia es negativa: se cumple cuando nadie interfiere ilegítimamente en la libertad del otro. Por eso no encuentra ningún problema con la ética de la justicia en el uso de aplicaciones de IA para la universalización de la salud: No hay deber de redistribuir datos, beneficios ni oportunidades. El papel del Estado debe limitarse a garantizar la seguridad, los contratos y la libertad de acción. Las desigualdades, mientras resulten de intercambios voluntarios, no constituyen una injusticia.

III.b.- La visión del mapa de las injusticias desde el punto de vista de Rawls

Para la teoría John Rawls, la justicia debe garantizar una sociedad equitativa. Para explicar como puede articular justicia y equidad formula dos principios: El primero afirma que cada persona debe tener un igual derecho al conjunto más amplio posible de libertades básicas compatibles con la libertad de los demás (como la libertad de pensamiento, de conciencia o de asociación). El segundo principio regula las desigualdades sociales y económicas, estableciendo que solo son justas si benefician a los más desfavorecidos y si se garantiza la igualdad de oportunidades.

Para poder definir como sería la aplicación de estos principios imagina una posición que denomina velo de ignorancia en la que los individuos, al diseñar las normas de la sociedad, desconocen su posición social, sus talentos o su estatus. De este modo, se eliminan los intereses particulares y se asegura que las reglas sean justas para todos.

1. Injusticia por pobreza de datos de salud

La ausencia de ciertos datos no es una omisión neutral. Es una forma de exclusión que actúa sin necesidad de declararse. Lo que no se registra, no se conoce, y lo que no se conoce no participa de las decisiones. En un sistema de salud que emplea inteligencia artificial, esta ausencia se traduce en tratamientos menos adecuados, diagnósticos incompletos, prioridades sesgadas.

Desde la perspectiva de Rawls, este fenómeno vulnera su primera premisa de justicia, que exige igual libertad y consideración para todos. Si determinados colectivos —minorías, mujeres, ancianos, pueblos pobres— no están presentes en los datos que alimentan los

algoritmos, entonces tampoco participan de las oportunidades que esos sistemas generan. El acceso equitativo a la salud comienza por ser representado en el conocimiento que la organiza.

Nadie, situado tras el velo de la ignorancia, aceptaría una arquitectura tecnológica que pudiera excluirle por defecto. La justicia, en este caso, no exige solo buenos resultados, sino condiciones equitativas desde el origen. Lo que se requiere es una redistribución epistémica que repare la estructura de los datos y garantice una representación justa.

2. Injusticia procedimental

Los algoritmos no solo producen resultados. También aplican reglas, toman decisiones y organizan prioridades. Cuando lo hacen de manera opaca, sin permitir la comprensión ni la deliberación, se rompe uno de los principios centrales de la justicia: la imparcialidad de los procedimientos.

Para Rawls, una sociedad justa debe poder explicar sus decisiones. No porque deba justificarse constantemente, sino porque la publicidad de los principios es una condición de su legitimidad. Un sistema cuya lógica permanece cerrada al ciudadano es una estructura sin rostro, en la que ya no es posible dialogar ni objetar.

3. Injusticia por exclusión de prioridades de investigación

Rawls establece que las desigualdades solo son aceptables si benefician a los menos favorecidos. Si la investigación científica se dirige hacia quienes ya tienen más acceso, más recursos o más representación, entonces la justicia ha sido desplazada por la conveniencia. La ciencia deja de ser una herramienta de equidad y se convierte en espejo de las desigualdades que ya existen.

Lo justo sería reorientar los recursos hacia quienes están en peor situación. No como acto de generosidad, sino como mandato institucional. La equidad exige, en este caso, una dirección consciente de la agenda científica.

4. Injusticia por falta de equidad en el reparto de beneficios

Cuando los frutos de la investigación no alcanzan a quienes han contribuido a hacerla posible, se rompe el principio de cooperación que sostiene la justicia. Comunidades que ofrecen datos, tejidos o colaboración no pueden quedar fuera del acceso a los beneficios que emergen de ese proceso.

Rawls defendería que los resultados de las aplicaciones de IA en salud deben favorecer prioritariamente a los menos aventajados. Y también que la equidad se expresa en la justa distribución de aquello que se genera en común. La justicia no se agota en respetar contratos; se expresa plenamente cuando el fruto compartido regresa a quienes más lo necesitan.

5.- Universalización de la salud, IA y el concepto de justicia de Rawls

La justicia, en la teoría de Rawls, se manifiesta allí donde las instituciones garantizan condiciones equitativas para todos, especialmente para quienes parten de posiciones más vulnerables. Las desigualdades no son neutras; deben ser examinadas a la luz de sus consecuencias para los más desfavorecidos.

Una inteligencia artificial justa, en este marco, no es aquella que funciona mejor, sino la que podría ser aceptada por cualquiera desde una posición de ignorancia sobre su propio lugar en el sistema. No se trata de eficiencia, sino de equidad moral.

III.- La visión comunitarista del mapa de las injusticias

La ética comunitarista entiende la justicia no como un conjunto de derechos individuales abstractos, sino como el bien común que surge de la vida compartida en comunidad. Frente al individualismo liberal, sostiene que las personas solo pueden comprenderse plenamente dentro de los vínculos, tradiciones y valores de la sociedad a la que pertenecen. La justicia, por tanto, no se define desde un punto de vista neutral o universal, sino desde los fines y virtudes que una comunidad reconoce como propios. Los comunitaristas — como Alasdair MacIntyre, Michael Sandel o Charles Taylor — afirman que una sociedad justa es aquella que fortalece los lazos sociales, fomenta la participación y orienta sus instituciones hacia el bien común, no solo hacia la protección de derechos individuales.

1. Injusticia por pobreza de datos de salud

Desde el comunitarismo, la justicia no se define por la imparcialidad, sino por la calidad de los vínculos que permiten que todos participen de un bien común. Cuando ciertos grupos no están representados en el conocimiento, no solo se debilita la precisión del sistema, sino el lazo que sostiene la comunidad.

Por tanto, para el comunitarismo, quedar fuera de los datos es también quedar fuera de la comunidad de cuidados. Allí donde la inteligencia artificial no registra, no hay mirada, y donde no hay mirada, no hay reconocimiento. No se trata solo de una ausencia técnica, sino de una exclusión silenciosa que refleja un déficit anterior de pertenencia.

2. Injusticia procedimental

Cuando las decisiones tecnológicas se toman de manera opaca, la comunidad pierde la posibilidad de deliberar sobre lo que le afecta. El algoritmo no responde, no explica, no escucha. En ese silencio, se rompe el diálogo y con él la posibilidad de construir decisiones compartidas.

La justicia, desde el comunitarismo, no consiste en aplicar reglas justas, sino en permitir que esas reglas nazcan del encuentro entre quienes comparten un destino común. Un sistema que no puede ser explicado no puede ser gobernado. Y lo que no puede ser gobernado, tampoco puede ser justo.

3. Injusticia por exclusión de prioridades de investigación

El comunitarismo defiende que el conocimiento debe responder a las necesidades reales de quienes forman parte de una comunidad humana. No basta con saber. Hay que saber para alguien. La orientación de la ciencia es también una expresión de su responsabilidad.

Por eso, cuando la investigación se orienta por la lógica del beneficio, se debilita el lazo moral que une a quienes investigan con quienes esperan ser beneficiados. No se trata solo de falta de acceso, sino de pérdida del sentido de comunidad. La ciencia deja de ser un proyecto común y se convierte en una empresa privada.

4. Injusticia por falta de equidad en el reparto de beneficios

Desde el comunitarismo, la justicia no se agota en un reparto equilibrado. Exige que lo que se genera en común fortalezca el vínculo que lo hizo posible. Compartir los beneficios no es una concesión, sino una forma de cuidar la cohesión social.

Por eso, quienes han contribuido a la construcción del conocimiento deben formar parte también de su fruto. Negar los beneficios a quienes han participado es romper la reciprocidad que hace posible la vida común. No se trata solo de proporción, sino de reconocimiento y de gratitud.

5.- Universalización de la salud, IA y el concepto comunitarista de Rawls

La justicia, en clave comunitarista, no se mide por la igualdad de los resultados ni por la transparencia de las reglas. Se mide por la calidad del vínculo que une a quienes comparten una vida en común. Una inteligencia artificial justa es aquella que reintegra a los excluidos, que permite deliberar en comunidad, que orienta su investigación al bien de todos y que distribuye sus frutos como expresión de un compromiso solidario.

Universalización de la atención médica gracias a la IA en la era de la hiper-especialización, Cuando la Filosofía ya no es suficiente⁶

1. Introducción. El sueño de la universalización

En este texto, me propongo adoptar un enfoque más general que el habitual al tratar el ideal de la universalización médica, tanto en los años previos al reciente auge de la inteligencia artificial (IA) como en la actualidad. Desde esta perspectiva –tradicionalmente denominada “filosófica”– intentaré fundamentar mi respuesta a la pregunta de si, con la IA, lograremos finalmente alcanzar la tan anhelada universalización o, al menos, avanzar significativamente en esa dirección.

Adelanto ya mi postura: considero que este ideal es factible a largo plazo; es decir, no lo percibo como una utopía. No obstante, sostengo que su realización no dependerá sustancialmente de la IA, sino de una gestión política adecuada, que a su vez requerirá un modelo educativo previo, sólido y comprometido. Por ello, es probable que debamos esperar –y quizá durante bastante tiempo– antes de observar avances significativos.

Asimismo, defenderé que, en este proceso, nos enfrentaremos a riesgos serios –reales, no distópicos– en los que, ahora sí, la IA podría desempeñar un papel causal determinante. Todos estos riesgos podrían condensarse en una misma imagen alegórica: la de una máquina insomne, incapaz de dormir y, por ello mismo, también incapaz de soñar.

Como provocación final, atribuiré un peso central al análisis histórico en mis reflexiones sobre el futuro tecnológico... pues el pasado, invariablemente, nos sorprende.

2. De la atención primaria a la Salud pública

⁶ Texto reelaborado a partir de una conferencia impartida como ponente invitado para el VIII Congreso sobre Derechos Humanos organizado por la Fundación Mainel. ‘Derechos Humanos e Inteligencia Artificial. La conferencia fue parte de la mesa redonda ‘Universalización de la atención médica gracias a la IA: ¿realidad, utopía o distopía? con la participación de Rafael Amo (Universidad de Comillas) y yo mismo, Luis Echarte (Universidad de Navarra). Valencia 16-17 de octubre de 2025.

La atención primaria representa uno de los principales ámbitos en los que la inteligencia artificial (IA) podría contribuir significativamente al ideal de la universalización. Para muchos, mejorar la digitalización y automatización de este servicio constituye una condición necesaria para reducir costes, monitorizar a los pacientes y personalizar los tratamientos (Sakar et al., 2024; Yip 2024). La propia Organización Mundial de la Salud (OMS) no solo ha ya reconocido el potencial de la inteligencia artificial en el ámbito de la salud pública, sino que está promoviendo activamente la elaboración de políticas en esta línea de trabajo (WHO, 2025).

Entre las dianas prioritarias señaladas en la literatura pueden destacarse cinco frentes principales de acción. Primero, el desarrollo de gemelos digitales de salud. Contar con un gemelo digital por paciente permitiría detectar de forma temprana alteraciones fisiológicas, calcular riesgos con mayor precisión y aplicar tratamientos más eficaces, oportunos y menos costosos. Segundo, la integración de asistentes médicos basados en IA, disponibles las 24 horas a través de dispositivos portátiles o domésticos. Estos sistemas podrían ofrecer diagnósticos inmediatos y mejorar significativamente la coordinación entre pacientes y profesionales humanos. Tercero, la transformación de los hospitales, que incluiría, entre otras innovaciones, el uso de robots quirúrgicos autónomos. Gracias a la inteligencia artificial, las intervenciones serían mínimamente invasivas, lo que reduciría los tiempos de recuperación y los riesgos asociados. Cuarto, la optimización de la gestión de los sistemas de salud pública. La IA permitiría una asignación más eficiente de los recursos, anticiparía brotes epidémicos mediante modelos predictivos a escala global y contribuiría a garantizar que nadie quedara excluido del acceso sanitario. Quinto, el establecimiento de mecanismos de gobernanza algorítmica que aseguren que las decisiones médicas automatizadas cumplan con los más altos estándares éticos. Esto implicaría avanzar hacia la eliminación progresiva de sesgos y la garantía de la equidad en el proceso de atención.

Como es lógico, estas cinco líneas estratégicas requieren que la atención primaria entre en diálogo con la salud pública, otro ámbito en el que la IA posee una proyección especialmente significativa. A su vez, es en este segundo nivel desde donde la medicina tiende sus puentes más sólidos hacia otras disciplinas como la economía, el derecho, la sociología, entre otras. Y es que, en efecto, si en algo coinciden los expertos es en que la universalización de la atención médica exige un abordaje profundamente interdisciplinar (Public Health, European Commission, 2025). Sin embargo, es justamente en este punto donde emerge el talón de Aquiles de la IA: su aplicación resulta mucho más sencilla y confiable en tareas concretas y delimitadas, mientras que su integración en contextos complejos, transversales e interdisciplinarios presenta importantes obstáculos (Panteli et al., 2025).

Sin embargo, pese a esta dificultad –o quizá precisamente a causa de ella–, debido a la indeterminación metodológica y conceptual que aún envuelve a la inteligencia artificial (IA), están proliferando grandes narrativas sumamente esperanzadoras sobre el brillante futuro que

nos aguarda: un porvenir en el que los servicios sanitarios se habrán transformado, por fin, en un proceso continuo de cuidado, educación y bienestar compartido.

Trataré de argumentar a continuación que este tipo de imaginarios no solo venden falsas promesas, sino que además pueden resultar extraordinariamente peligrosos.

3. Un nuevo ELIZA

La utopía científica tiene sus propios ritmos, que suelen acompañarse con descubrimientos y avances tecnológicos significativos. Así ocurrió en los años sesenta con el surgimiento de la inteligencia artificial. Fue durante esa década cuando se desarrollaron los primeros programas y máquinas capaces de simular tareas cognitivas humanas. Uno de los casos más emblemáticos fue ELIZA, el célebre programa creado por Joseph Weizenbaum en 1966, que simulaba una conversación con un terapeuta: el primer chatbot de la historia. Algunos llegaron a augurar que el acceso universal a la salud mental estaba a la vuelta de la esquina, disponible para cualquiera que contara con un ordenador personal en casa. Los ordenadores personales llegaron, efectivamente, en la década de los setenta. La salud mental, en cambio, no. Para algunos, incluso, llegó a deteriorarse con el uso de dichos dispositivos.

Pues bien, la historia se repite, aunque ahora las promesas tienen un alcance aún mayor. En los foros de bioética contemporáneos se debate sobre la posibilidad de la supercognición, la superlongevidad o el superbienestar: narrativas que, en muchos casos, también integran a la inteligencia artificial (IA) como parte central de sus propuestas.

Un caso paradigmático lo ofrece *The Age of Scientific Wellness* (2023), de Leroy Hood y Nathan Price, donde no solo se plantea cómo la IA puede fomentar la salud, sino que se llega a afirmar que es capaz incluso de alcanzar la felicidad humana, rebautizada bajo el concepto de “bienestar científico”. Y aquí empiezan los problemas: no en el intento –legítimo y necesario– de vincular salud y felicidad, sino en la reducción de la segunda a la primera; es decir, en identificar la felicidad, como bien general, con la salud, que hasta hace poco era considerada un bien particular.

En esta concepción tecnificada de la felicidad, la IA asume un protagonismo absoluto, erigiéndose en herramienta clave para pasar de la teoría a la práctica: un logro tecnológico que, sin embargo, puede derivar en un malogramiento de lo más específicamente humano. Veamos ahora de cuántas formas distintas los avances de la IA pueden afectar negativamente al progreso, tanto individual como social.

4. Distopías médicas

Aunque no son hoy tan numerosas como las narrativas tecno-optimistas, las alertas y distopías tecnológicas también pueblan los debates contemporáneos, ya sea en forma de

discurso académico –como evidencian las numerosas críticas dirigidas al transhumanismo– o en clave literaria.

Entre estas últimas, dos ejemplos recientes y destacados son *Klara y el Sol*, del premio Nobel de Literatura Kazuo Ishiguro (2021), y *Máquinas como yo*, de Ian McEwan (2019). Ambas obras plantean, desde enfoques distintos, escenarios en los que la inteligencia artificial (IA) se convierte en la principal fuente de conflictos globales. Veamos algunos lugares comunes que comparten estas narrativas distópicas.

En primer lugar, se advierte el riesgo de que, al estar cada ciudadano constantemente monitorizado por sensores biométricos, se genere un flujo continuo de datos que llegue a determinar su valor sanitario y, en consecuencia, su acceso a tratamientos médicos.

En segundo lugar, este peligro se vería agravado por la existencia de sistemas de IA centralizados, gestionados por corporaciones o gobiernos con intereses partidistas. Serían ellos quienes decidirían quién merece atención prioritaria, según criterios como la productividad, el historial genético o el costo estimado para el sistema. Tales decisiones, además, probablemente terminarían delegándose a la propia inteligencia artificial.

En tercer lugar, si se avanzara en esta dirección, sería previsible que ambulatorios y hospitales quedaran reducidos a centros casi completamente automatizados, con mínima presencia de personal médico. En estos espacios se atendería únicamente –como se acaba de plantear– a los pacientes aprobados por el algoritmo, mientras que los demás recibirían cuidados remotos... y de baja calidad.

En cuarto lugar, la relación médico-paciente sería reemplazada por simuladores afectivos: interfaces emocionalmente vacías, pero altamente persuasivas. De este modo, ni siquiera en ese último bastión de lo humano encontraría el profesional sanitario un refugio, quedando aún más atrapado –y más frustrado– en tareas técnicas y administrativas.

Por último, en quinto lugar, no solo se habría reducido el derecho a la atención médica a un privilegio algorítmico, sino que el control de la salud se convertiría en un mecanismo de vigilancia social. Los datos biométricos podrían emplearse para rastrear conductas, emociones y desviaciones respecto del *bienestar normativo*.

5. Tecnomedicina postmoderna

El principal interés de los escenarios irreales descritos en el epígrafe anterior radica en que permiten visibilizar riesgos que solo se comprenden plenamente cuando se adopta una mirada amplia y contextual. Y, como luego explicaré, no es casual que la bioética contemporánea adolezca, en ocasiones, de los males del cortoplacismo y del particularismo.

Con todo, hay quienes se resisten a los males de época. Entre ellos, destacan aquellos autores que van más allá de la especulación distópica y sitúan sus críticas en el terreno áspero —y muy real— del presente.

Por ejemplo, especialistas en inteligencia artificial como Geoffrey Hinton y Yoshua Bengio, ambos galardonados con el prestigioso Premio Turing, llevan tiempo alertando sobre los riesgos asociados al mal uso de esta tecnología.

Sin embargo, en el ámbito biomédico resultan especialmente relevantes las investigaciones de José Repullo, quien durante décadas ha sido profesor de Planificación y Economía de la Salud en la Escuela Nacional de Sanidad, así como investigador del Instituto de Salud Carlos III. Todo un referente en España en materia de Salud Pública.

Pues bien, al profesor Repullo le preocupa la creencia —cada vez más extendida— de que la tecnología pueda erigirse como el principal remedio para todos los problemas sanitarios. Una idea que, según advierte, está concentrando los principales focos de inversión y orientando las líneas de acción de la política sanitaria actual —no solo en España.

Para Repullo si no ponemos remedio, con la tecnomedicina post-moderna, clínicos y paciente acabarán adaptándose y acoplándose a la máquina y no al revés. De hecho, a su juicio, la tecnología ya ha empezado a “redefinir el contorno del acto médico, y a disciplinar a todos en torno suyo” (Repullo 2024). Es decir, puede que la universalización se acabe produciendo, pero ¿a qué precio? El de una medicina cada vez más fragmentada y atomizada y, por ello, finalmente rendida a la inmediatez clientelar y la mercantilización comercial. Esto la consecuencia, concluye Repullo, de confundir los fines y los medios: primero, la tecnología con la salud y, segundo, la salud con la felicidad.

Por último, José Repullo divide la deriva del modelo sanitario hacia su polo más tecnificado y tecnificante en tres etapas: en la primera fase, el centro de gravedad se sitúa en la relación médico-paciente, siendo la tecnología un recurso esencial, aunque no definitorio; es decir, el sentido del acto médico no depende de ella.

En la segunda fase, en la que nos encontraríamos actualmente, ese centro se desplaza hacia los equipos asistenciales. En este estadio, la tecnología aún no es determinante, pero empieza a influir de manera creciente en la comprensión y mediación de la relación médico-paciente.

En la tercera y última fase, el protagonismo lo asumirían las baterías diagnósticas y terapéuticas, que escalarían al primer plano. Aquí sí, la tecnología se consolidaría como la principal fuente de legitimidad en la toma de decisiones asistenciales —signifique esto lo que pueda significar.

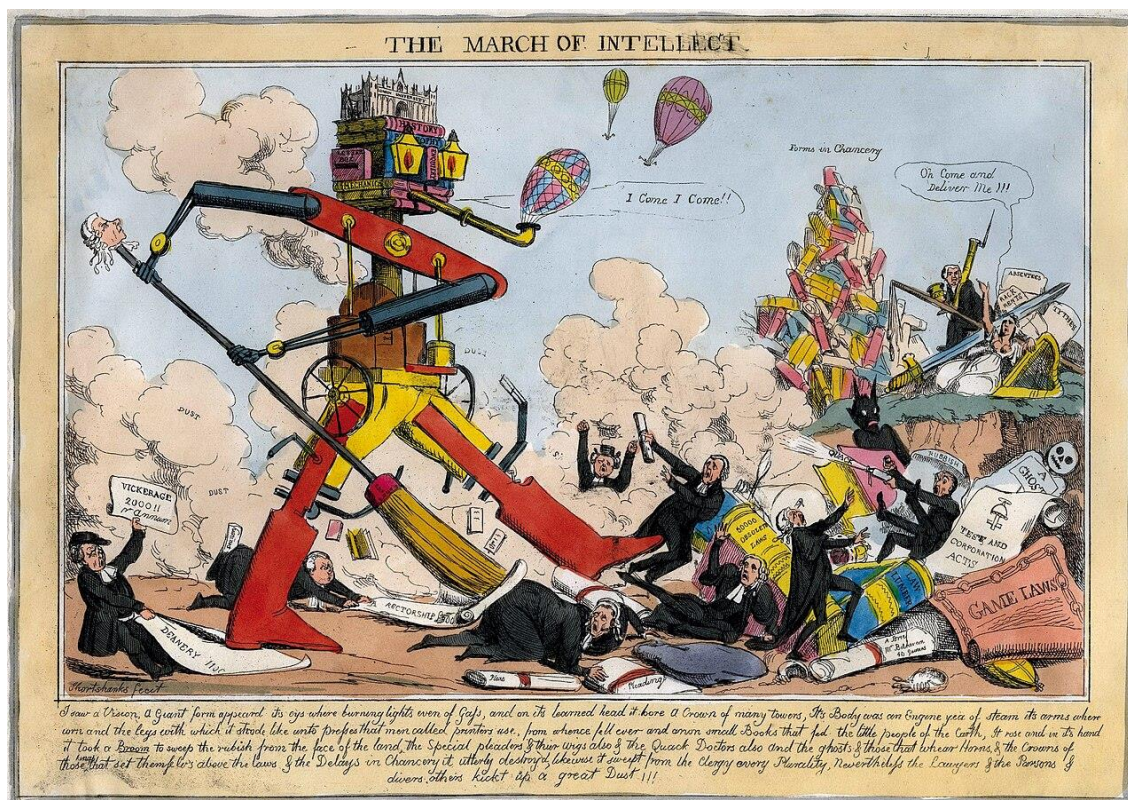
6. La pobre fama de los augures

Ante el diagnóstico sistémico de Repullo, caben al menos dos preguntas fundamentales. En primer lugar: ¿cómo hemos llegado hasta aquí? Y en segundo: ¿qué acogida están teniendo estas llamadas de atención? Comencemos por esta última.

Pese al morbo que suelen despertar las argumentaciones catastrofistas, su impacto real en los espacios de decisión es, por lo general, bastante limitado. La distopía funciona bien en la literatura y en el cine; en cambio, la utopía tecnológica suele ser mucho mejor recibida en los foros científicos y en los comités de expertos... que son, con suerte, los lugares donde se definen las políticas sanitarias. Y esto, en el fondo, tiene algo de profundamente humano: por un lado, tendemos a responder con mayor apertura a los refuerzos positivos que negativos; y por otro, somos —por lo general— más indulgentes con las falsas esperanzas que con los mensajes agoreros sobre futuros apocalípticos.

Pero hay más razones, y estas tienen que ver con el espíritu de nuestra época; primero, tras las cuatro grandes revoluciones tecnológicas de los últimos dos siglos —siendo las dos más recientes la de las comunicaciones y la de la inteligencia artificial—, hemos comenzado a acostumbrarnos, valga la paradoja, a que la tecnología nos sorprenda. Esto se traduce en una creciente desconfianza hacia la idea misma de imposibles tecnológicos. A este fenómeno se suma otro, de carácter ideológico, vinculado a la persistencia del dogma del progreso tecnológico inevitable. Me refiero a la vieja noción del humanismo ilustrado, según la cual el avance de la tecnología constituye un proceso lineal e inexorable, que conduce de forma natural al mejoramiento social, económico y cultural —de ahí su arraigado tecno-optimismo.

El asunto viene de largo. Para referirme a éste me gusta poner de ejemplo *La marcha del intelecto*, una caricatura de Robert Seymour, publicada en 1828, donde el espíritu ilustrado toma, no por casualidad, forma de máquina barriendo a políticos, filósofos, artistas y pastores... Es el sueño ilustrado de una técnica, por fin, poniendo orden en todos los asuntos humanos.



El dibujo es ya de dominio público y el original se encuentra expuesto en el Museo Británico, junto a otras obras vinculadas a la *Marcha del Intelecto*. Este fue el nombre que recibió una controversia que provocó un encendido debate en la Inglaterra de comienzos del siglo XIX. Mientras algunos sectores la celebraban como un emblema del avance social hacia una mayor difusión del conocimiento, la comprensión y finalmente la paz y la justicia social, otros la denunciaban como una muestra de la obsesión moderna por el progreso y la innovación.

Conviene recordar que esta disputa se produjo en un contexto especialmente convulso para la sociedad británica, atravesada por procesos de industrialización acelerada, el auge de ideales democráticos y profundos cambios en las estructuras sociales tradicionales. Y en efecto, en lo que respecta a las transformaciones sociales, quizá no vivimos en un tiempo tan distinto de aquel.

7. Frente a los dioses castigadores

Para los ilustrados —de ayer y de hoy—, la razón matematizante y sus frutos tecnológicos representan el arma definitiva con la que conquistar el Olimpo: derrotar a los antiguos dioses y ocupar su lugar.

Por supuesto, sería erróneo afirmar que esta idea sea propiamente ilustrada. Confrontemos, la caricatura satírica de Robert Seymour con una de las grandes obras de Pieter

Paul Rubens: Prometeo llevando el fuego a la humanidad (1636), obra que hoy se conserva en el Museo del Prado.



Es una pintura de profunda fuerza pictórica y simbólica, que remite a uno de los mitos más potentes del mundo clásico. Prometeo es un titán que se rebela contra los dioses del Olimpo para ayudar a la humanidad. Compadecido por la condición frágil y vulnerable de los seres humanos, robó el fuego sagrado de los dioses— símbolo del conocimiento, la técnica y el poder divino— y lo entregó a los hombres para que pudieran progresar y desarrollarse.

Me gusta acompañar la obra de Rubens con unos versos que Goethe dedicó también a Prometeo, y que expresan con gran fuerza poética esa nueva esperanza propia de la escatología laica. Entre 1772 y 1774, escribe:

Aquí estoy, dando forma [canta Prometeo retando a Zeus]
a una raza según mi propia imagen,
a unos hombres que, iguales a mí, sufran
y se alegren, conozcan los placeres y el llanto,
y, sobre todo, a ti no se sometan,
como yo [como yo, a disgusto, me someto a ti, oh Zeus].

En estos versos, el poeta ya vislumbra la larga sombra que proyecta el faro de la Ilustración, y que más tarde llegará a criticar abiertamente. Sin embargo, Goethe también es

consciente de que, frente a semejante promesa de prosperidad y autodeterminación, no solo el cuestionamiento, sino la mera duda, constituye ya una auténtica afrenta al espíritu del hombre moderno.

¿Quién está realmente dispuesto a cargar con el estigma de un pensamiento tachado de retrógrado —y que, además, suele asociarse al discurso del miedo, del autoritarismo y, por supuesto, al severo juicio de los dioses castigadores? Por el contrario, en el horizonte de la esperanza ilustrada, el fuego no desciende del cielo: el fuego asciende, viene de abajo hacia arriba, y lo hace siempre para bien.

8. Si monumentum requiris...

Aldous Huxley es uno de los intelectuales del siglo XX que más páginas dedicó a estudiar y denunciar el prejuicio objetivista, el tecno-optimismo y lo que hoy podríamos denominar cultura de la cancelación. Veamos algunas de sus ideas más relevantes para evaluar las consecuencias del reduccionismo tecno-optimista.

Lo primero que conviene señalar es que Huxley hablaba con pleno conocimiento de causa. Su obra más célebre, *Un mundo feliz* (1932), fue recibida con entusiasmo por parte del público, pero también suscitó las más feroces críticas entre los científicos y filósofos de su época. Sin embargo, el tiempo acabó dándole la razón: *Un mundo feliz* es una novela visionaria que anticipó con asombrosa precisión la revolución tecno-cultural de la segunda mitad del siglo XX.

No es de extrañar, entonces, que, en el prólogo a la edición de 1946, Huxley se permitiera su particular revancha. Huxley recuerda con sarcasmo cómo su novela fue, en su momento, tachada como “un síntoma del fracaso de una clase intelectual en tiempos de crisis”. Y esto se lo decían precisamente aquellos que se consideraban a sí mismos como “los principales representantes del éxito y del progreso moderno”. Pues bien, a todos ellos Huxley les dedica las siguientes palabras —implacables—:

[Estos] benefactores de la humanidad se merecen el debido honor y conmemoración. Construyamos un Panteón para académicos. Deberíamos enclavarlo entre las ruinas de las ciudades destruidas de Europa o Japón, y sobre la entrada del osario yo grabaría, en letras de unos dos metros de altura, estas simples palabras: «consagrado a la memoria de los educadores del mundo.

La cita termina con el siguiente adagio latino:

SI MONUMENTUM REQUIRIS, CIRCUMSPICE

Al igual que José Repullo, Huxley sostiene que hay ideas extremadamente peligrosas, pero va más allá. Termina por relacionar los peores horrores del siglo XX con la creencia de que los grandes sueños de la humanidad —la salud, la riqueza, la educación— pueden alcanzarse mediante un método puramente reductivo. Una creencia que no solo es falsa, sino que, además, desemboca en las formas más extremas de horror y deshumanización. Porque, se pregunta Huxley con contundencia, ¿Quién puede seguir afirmando que la educación ilustrada previene al ser humano de cometer actos de barbarie como son el lanzamiento de armas nucleares sobre miles de niños, mujeres y ancianos?

Los abordajes parciales solo resuelven problemas locales. *Un mundo feliz* constituye, bajo dicha luz, una excelente narrativa sobre los efectos del proceso de tecnificación de lo humano, que convierte al cuerpo en una fuente inagotable de control, y a la medicina, en un instrumento de sumisión. Ahora bien, Huxley también se esforzó por dejar claro que no estaba defendiendo una postura anti-tecnológica —aunque no tuvo demasiado éxito en librarse de ese sambenito. Precisamente por ello dedicaría los últimos años de su vida a la redacción de *La isla* (1962), una novela verdaderamente utópica —o mejor dicho, de un realismo optimista— en la que la educación sigue ocupando un lugar central, aunque ahora para bien.

Dedicaré la siguiente sección a extraer los puntos más interesantes de una novela que se publicaría un año antes de la muerte de su autor.

9. Universalización sin alta innovación

En *La isla*, Aldous Huxley expone su visión de lo que debería ser una verdadera formación integral, tanto en el plano intelectual como en el emocional y sensorial. Una formación que, por tanto, no se orienta al adoctrinamiento ni al control, sino al desarrollo pleno de la conciencia humana.

En esta novela, Huxley retoma la misma estrategia narrativa que utilizó en *Un mundo feliz*: construir una ficción que, sin contar con pruebas empíricas sobre el futuro, tampoco quede limitado por la ausencia de evidencias. Frente a la incertidumbre, compensa la falta de datos con la coherencia interna del relato. En otras palabras, opta por un enfoque radicalmente interdisciplinar, en el que confluyen no solo dimensiones científicas o históricas, sino también —y esto es clave— estéticas.

Desde esta perspectiva, *La isla* representa una propuesta audaz: imaginar un futuro en el que se trabaje al mismo tiempo la totalidad y la profundidad, lo objetivo y lo subjetivo, lo externo y lo íntimo. Una utopía que se aleja del reduccionismo tecnocrático para apostar por un equilibrio integral entre conocimiento, sensibilidad y espiritualidad. Los ciudadanos de Pala, la sociedad imaginada por Huxley en *La isla*, aprenden a reconocer el lugar de cada fin y de cada medio en una imagen general de mundo, incluida la tecnología, que lo permea todo, hasta

la propia espiritualidad, sin que de ello se derive la confusión entre los fines y los medios. Sobre esto escribirá: “No existe una sola cura para lo que no puede tener una sola causa [...] Nada sirve, como no sea el todo” (Huxley 2003, 183).

En dicha isla, la atención médica es verdaderamente universal, pero no porque dependa de la tecnología más avanzada, sino porque sus ciudadanos han sido educados en el valor de la comunidad y en el respeto por los derechos humanos. Y no solo eso: también han sido prevenidos contra los peligros de los “ismos” —el hedonismo, el consumismo, el populismo o el egoísmo—, esas ideologías y hábitos que, para Huxley, funcionan como distracciones modernas que bloquean una reflexión seria y profunda sobre la relación entre salud, éxito y felicidad.

Como consecuencia de esta educación, y solo a largo plazo, llegan al poder líderes verdaderamente sabios: una clase política capaz de gobernar con discernimiento, de distinguir entre fines y medios, y de distribuir con justicia tanto los bienes materiales como los intelectuales. Huxley es plenamente consciente de que ni siquiera la mejor educación puede garantizar la llegada de los mejores líderes, pero también sostiene que es condición necesaria para facilitar que eso suceda.

Finalmente, aunque no ocupa el lugar central, la tecnología sí está presente en todas las esferas de lo humano, incluidas las más elevadas. Es en este contexto donde aparece, por ejemplo, el uso ritual y terapéutico de la *medicina Moksha*, una sustancia imaginada que ayudaría a que el individuo estableciera mejores puentes entre lo material y lo espiritual.

10. El polvorín de Sagan

La isla ejerció un impacto notable en la cultura de su época, pero no ha perdurado en el imaginario colectivo con la misma fuerza que *Un mundo feliz*, obra que sigue siendo citada con frecuencia en los debates contemporáneos sobre bioética.

Esto se debe, en parte, a que muchas de las profecías formuladas en *La isla* aún están en proceso de cumplirse —lo que mantiene viva su vigencia crítica—, y también a que *La isla* suele ser percibida hoy como una propuesta algo vaga o ingenua. Sus ideas no conducen fácilmente a planes estratégicos concretos ni a acciones directas susceptibles de evaluación empírica, lo que dificulta su integración en los marcos técnicos y operativos actuales.

Sin embargo, esa primera valoración, ese juicio moral inmediato o simple palpito, no es trivial: es síntoma del problema de fondo. Tiene que ver con cómo hemos sido educados... en el gusto por la compartimentalización y en la impaciencia por obtener logros a corto plazo. Así lo reconoce el célebre astrónomo y divulgador Carl Sagan en una entrevista concedida a Charlie Rose en 1996. “Hemos organizado una sociedad basada en la ciencia y la tecnología.” Y justamente eso, lo que Sagan enuncia con orgullo, es lo que haría palidecer a Huxley. Ahora bien, esa discrepancia no impide que ambos compartan una preocupación común, que Sagan

expresa con particular claridad en la continuación de la misma entrevista: “Hoy nadie entiende nada de ciencia y tecnología. Y esta mezcla explosiva de ignorancia y poder, tarde o temprano, terminará explotándonos en la cara. Porque..., ¿quién está dirigiendo la ciencia y la tecnología en una democracia donde el pueblo no entiende nada sobre eso?” (Sagan 1996).

Sobre este último punto, Huxley replicaría a Sagan que el problema que denuncia es inherente a la misma organización social que Sagan celebra con tanto orgullo. Porque una sociedad estructurada bajo criterios puramente científicos y tecnocráticos solo puede dar lugar a una cultura de especialistas. Y desde la pura especialización, advierte Huxley, no es posible comprender qué es la ciencia en su totalidad, y mucho menos imaginar qué puede haber más allá de la razón matematizante o, en el contexto actual de la inteligencia artificial, qué queda fuera del pensamiento algorítmico.

Desgraciadamente, no habitamos en la isla de Huxley, sino en el polvorín de Sagan: la sociedad en la que los ciudadanos caen en la ingenuidad de preguntarse si con el auge de la IA se logrará, por fin, la tan anhelada universalización de la atención médica. Peor aún: la nuestra es una sociedad que espera que la implementación de la IA en las aulas sea capaz de mejorar la educación y extenderla a todos los rincones del planeta, sin importar nacionalidad, cultura o situación económica. Para Huxley, esto sería la abominación de la desolación pues, lo que realmente conseguiremos, configurando la educación en torno a la IA —y no al revés—, será el surgimiento de escuelas y universidades aún más especializantes, y por ello mismo, más embrutecedoras y homogeneizantes. Más leña al fuego, en este ya peligroso polvorín de ignorancia y poder.

11. Dos tipos de fuego amigo

Cambiemos de tercio. Pasemos ahora a analizar el fenómeno del fuego cruzado, del fuego amigo —ya se sabe—, a veces más letal que el que dispara el enemigo.

En los años cincuenta del siglo pasado, Aldous Huxley celebra los nuevos impulsos sociales que emergen en reacción al proceso de tecnificación social —no olvidemos que es la década de la *generación beat*. Sin embargo, al mismo tiempo, detecta no pocos lobos con piel de cordero, especialmente en el ámbito académico. Se trata de aquellos que, por un lado, defienden a ultranza el progreso tecnológico, para luego matizar —con cierta condescendencia— que sí, existen riesgos, pero que estos podrán ser neutralizados con simples campañas de concienciación, eso sí, anunciadas a bombo y platillo. La historia, también en este punto, tiende a repetirse: “lavado ético” es como hoy muchos se refieren a este mismo fenómeno de racionalización moral.

Pero Huxley no solo se refiere a quienes hacen trampa conscientemente. También señala a aquellos que, aun bienintencionados, perciben el problema, pero no su verdadera

magnitud. Por eso las iniciativas que se proponen en este segundo grupo no representan más que meros barnices, capas protectoras que solo se ocupan de lo superficial pero no del problema de fondo. Es lo que denuncia en *La isla*, cuando escribe: “La cura que tienen para el exceso de especialización científica consiste en unos cuantos cursos más de humanidades.” Y añade, con ironía: “No son más que otra forma de especialización” (Huxley 2003, 291). La propia filosofía ha enfermado de especialismo, se queja el novelista. Los médicos que podían curarnos han comenzado a enfermar también, y ya no salvan, sino que se han transformado en el peor vector de contagio. ¿La razón? Para Huxley, la prudencia y la humildad que les caracterizaba se ha vuelto pose, un delgado paño que apenas sirve ya para cubrir las vergüenzas. Pero lejos de mofarnos de ver al rey desnudo rodeado de su corte, afirma Huxley, debiéramos temblar, porque nos va en mucho su desgracia.

Esta última crítica sigue siendo plenamente relevante en nuestro presente. Somos testigos hoy de medidas promovidas por un humanismo superficial, de boquilla, o como a mí me gusta llamarlo: un “humanismo cool”, que pretende nadar y guardar la ropa; sumarse a la cruzada de las grandes esperanzas objetivistas, pero sin uniforme ni insignias, por tanto, sin la vergüenza que les causaría vestir el hábito del cientificismo.

Como es obvio, Huxley no cae en el error de los extremos. No rechaza ni la tecnología ni la formación especializada, sino que las sitúa en el lugar donde mejor pueden contribuir al verdadero progreso humano. Sus palabras no dejan lugar a duda: “La especialización en ese sentido es necesaria e inevitable. Sin especialización no hay civilización. Y si uno educa toda la mente-cuerpo junto con el intelecto utilizador de símbolos, ese tipo de especialización necesaria no produce mucho daño” (Huxley 2003, 291).

Lamentablemente, los modelos educativos actuales parecen avanzar en la dirección opuesta, o al menos, pretenden invertir el orden de los factores: partir de la formación especializada para alcanzar una formación integral, concebida como la mera suma de las partes. La propuesta huxleriana de que hay que empezar por el todo hoy se nos antoja locura.

12. No temer nada y temerlo todo

Huxley no propone –insisto– una renuncia a la tecnología ni un regreso nostálgico al pasado, sino una auténtica rehabilitación del humanismo, eso sí, introduciendo novedades significativas. Podemos llamarla, como se hace desde hace algunos años, humanismo tecnológico. Huxley, sin embargo, prefiere hablar de un retorno a la filosofía perenne.

Aquí emplearé el primer término, con el fin de sintonizar con la terminología de mis contemporáneos. Examinemos, a continuación, las principales características del humanismo tecnológico que encontramos en *La Isla*.

1. Tecnología como medio, no como fin

El humanismo tecnológico parte de una premisa esencial: reconocer el enorme poder de la tecnología, y al mismo tiempo, no perder de vista que sigue siendo solo un medio. Esta doble aproximación comprensiva –valorar sin idolatrar– no es fácil de lograr, pero es fundamental si queremos evitar tanto el rechazo ciego como la rendición acrítica.

2. Recuperación y actualización del viejo humanismo

El humanismo tecnológico recupera y resignifica muchas tesis del humanismo clásico, que sigue albergando tesoros intelectuales y éticos, algunos ya descubiertos pero olvidados, otros aún por descubrir.

Un buen ejemplo es la preocupación renacentista por una educación integral, capaz de forjar un nuevo sentido común colectivo. Frente a los éxitos del especialismo moderno, el viejo humanismo reivindica la solo aparente mediocridad del diletantismo: saber un poco de todo no es defecto, sino puerta al juicio complejo y a la sensibilidad plural.

La idea también está presente en *La Isla*. Huxley describe la superioridad palanesa como un arte que “si bien más elevado y mucho más valioso que todos los demás, puede ser practicado por todos: el arte de experimentar en forma adecuada, [...] el arte de conocer de modo más íntimo todos los mundos que, como seres humanos, estamos habitando” (Huxley 2003, 240).

Que nadie se confunda: de ese caldo cultural, advierte Huxley, surgieron los grandes genios del Renacimiento. Toda una lección para la nueva escuela del no-liderazgo y del no-éxito. Quizá vivimos demasiado obsesionados con alcanzar el Premio Nobel de la Paz... y así nos va. Pero que nadie se sienta excluido: podría decirse lo mismo del Premio Nobel de Medicina, de una cátedra universitaria o de un Óscar. Da igual. Para Huxley, esta hoguera de vanidades, esta *rat race*, suele justificarse en nombre de los ideales más elevados. De nuevo, el sesgo de la racionalización no parece el mejor método para construir progreso.

3. Nuevas preguntas para una nueva condición humana

El humanismo tecnológico no se limita a actualizar viejos principios: se atreve a formular preguntas nuevas. Por ejemplo: ¿quién es ese animal que posee la capacidad de modificar su propia naturaleza? ¿O incluso de autodestruirse –no solo como individuo, sino como especie?

Esta es una cuestión inédita en la historia del pensamiento humano, surgida del encuentro con las nuevas tecnologías, especialmente las de escala planetaria.

4. La virtud doble: esperanza y miedo

Finalmente, Huxley subraya una doble virtud que hoy suena políticamente incorrecta, pero que resulta indispensable: enseñar a los jóvenes a esperar lo mejor del ser humano –la virtud de la esperanza–, pero también a temer lo peor –la virtud del miedo.

Huxley prefiere hablar de la virtud del valor, porque considera que ésta segunda no puede forjarse sin pasar por la experiencia del miedo. El miedo, para él, no es intrínsecamente negativo: informa de los riesgos y les otorga significado. El peligro abstracto apenas despierta conciencia; el peligro temido, en cambio, sí lo hace. Después, una vez comprendidas las causas del miedo, éste puede ser eliminado. Porque vivir en un estado constante de temor tampoco es deseable ni saludable.

Tampoco Huxley está diciendo nada nuevo en este sentido. La idea ya aparece en *Las Leyes* de Platón, donde el filósofo propone, precisamente, un modelo de ciudad cuya prioridad sea educar a los ciudadanos... como ciudadanos. Ése es el contexto en el que Platón advierte contra una educación que aisle a los jóvenes del miedo o, al menos, de los grandes miedos. Entonces se hace inevitable contraer el vicio de la temeridad.

“¿Pero no tenemos un brebaje, cuyo efecto es inspirar una seguridad y una confianza temerarias e indebidas [...] llenando su alma de mil bellas esperanzas; dándole una idea más ventajosa de su poder, y, por último, inspirándole una plena seguridad de hablar de todo como si nada ignorara, y haciéndole de tal manera libre, de tal manera superior a todo temor, que sin detenerse, dice y hace todo lo que le viene á la imaginación?

El filósofo concluye con una recomendación que trasciende todo lugar y tiempo. En fin, un auténtico elogio del verdadero valor.

“Hay dos cosas, en las que es preciso aguerir nuestra alma; la una, no temer nada es ciertas ocasiones; y la otra, temerlo todo en otras. [...] “Es preciso, pues, que cada uno de nosotros no tenga miedo y sea temeroso a la vez” (Libro I, *Las leyes*, Platón).

13. Otra lección sobre el no-liderazgo

Hasta ahora he argumentado que, si buscamos la universalización de la atención médica priorizando soluciones tecnológicas, corremos el riesgo de deshumanizar el sistema hasta el punto de causar daños globales irreversibles. Además, podríamos estar entregando a nuestros gobernantes una herramienta sin precedentes, cuyo poder sería muy difícil de revocar en el futuro. No sería ninguna novedad: la historia demuestra que, cada cierto tiempo, el poder vuelve a concentrarse en figuras autoritarias. Siempre hay un nuevo Calígula esperando su turno. Lo que puede cambiar es el contexto, la singularidad tecnológica del poder omnívoro en las peores manos.

Estas son las dos razones principales por las que debemos devolver a la educación el lugar central que le corresponde, por encima de la innovación tecnológica, y también por encima del marco legal. Porque si la aplicación de la ley depende del juicio ético, este, a su vez, se sustenta en una educación integral. Esta y no otra es por la que Sócrates se considerara, ante todo, un educador. No solo de ideas, sino también de emociones, de carácter. El pensamiento socrático es un revulsivo idóneo para emprender el camino de la desespecialización filosófica, para devolver la filosofía a la raíz de su actividad, que no es la parte, sino el todo y el fondo. Sócrates promueve la educación en un sentido muy concreto, que es el mismo que un buen padre desea para sus hijos: la educación en la paideia, esto es, en esos ideales compartidos y eternos que constituyen la naturaleza humana, nunca terminada por descubrir, y que su propia adquisición y aplicación son fines en sí mismos, más allá de los resultados de la acción.

Pero, ¿están dispuestos los filósofos actuales a volver a reconocer que no saben? Y aún más: ¿a reconocer que, quizás, sean quienes menos saben entre todos los profesionales? Una segunda cuestión, no menos importante, es si podemos esperar dicha actitud socrática entre los filósofos de mayor éxito. Por motivos obvios, les resultará más difícil ese despojamiento intelectual que a los que no han logrado escalar en la carrera académica. A esto se refiere, por ejemplo, el filósofo suizo Alain de Botton, cuando retoma la célebre frase de Montaigne — “filosofar es aprender a morir” — y la imagina pronunciada por un estudiante o un profesor del Departamento de Filosofía del King's College de Londres. “Te mostrarán la puerta de salida. No es eso lo que sus miembros creen que están haciendo” (Botton, 2025).

Al contrario, resulta más plausible imaginar que la actitud de la vieja filosofía pueda emerger en una universidad de segunda fila o entre profesores de perfil bajo, más libres —y acaso más dispuestos— a conspirar contra un modelo universitario que, como ha advertido el profesor Alejandro Vigo, sufre una profunda tecnificación en todas sus áreas, y especialmente en la de la filosofía (Vigo, 2022). Este fenómeno también explicaría el distanciamiento de la filosofía respecto a los grandes asuntos, y particularmente respecto al más importante de todos: el de la felicidad. “Si quiere vivir mejor, enamórese, tome Prozac, no busque en la filosofía.” (Lipovetsky, 2025). Son palabras provocativas de uno de los filósofos con más visibilidad internacional en las últimas décadas y que recogen un sentir amplio académico. Y es lógico que la filosofía no de la filosofía si ya no consiste en lo que consistía. ¿Y qué consiste hoy? Para Slavoj Žižek, otro de los filósofos superstar de nuestro tiempo, la filosofía debe dejar abandonar el escrutinio por la realidad y, en particular, por la realidad de uno mismo al que llegamos por autoexamen y diálogo —la anagnórisis. En otras palabras. Sócrates se equivocaba. Tenemos que olvidarnos de nosotros mismos porque en las profundidades del yo solo habitan monstruos. Es mejor fingir lo cada uno quiera ser (Žižek 2002, 2006). ¿Filosofía de éxito o antifilosofía?

14. Lo que verdaderamente ha venido para quedarse

Volvamos a la antigua idea de filosofía que, bajo la luz socrática, no consiste en otro tipo de ocupación más, sino aquella que permite realizar cualquier actividad de la manera más elevada. De hecho, la institución universitaria se caracterizó durante siglos por este afán: entregar a la sociedad médicos filósofos, abogados filósofos, arquitectos filósofos... todos doctores en cuanto filósofos.

¿Podremos salvar la institución o habrá que sacar la filosofía de la academia para su revitalización? El filósofo surcoreano Byung-Chul Han, profesor de la Universidad de las Artes de Berlín, espera que no sea necesario que ocurra tal desastre, aunque, sin embargo, sí que anticipa que no será posible sin un terremoto previo. En una rueda de prensa concedida a los medios antes de recoger el premio *Princesa de Asturias de Comunicación y Humanidades* llegó a afirmar que: “Los catedráticos son vendedores, los estudiantes son clientes que acaban evaluando a los profesores. Este es el colapso de la cultura. Y mi tarea es mostrar ese colapso [...] Tengo la esperanza de que colapse el sistema, y va a pasar pronto” (Han 2025).

El asunto de la supervivencia de la institución universitaria es secundario al problema principal: sin un plan educativo sólido –se implante desde donde se implante–no solo será imposible alcanzar la universalización de la atención médica, sino que tampoco podremos prevenir los grandes riesgos asociados a la inteligencia artificial. Ninguna regulación, ley o filosofía serán suficientes. Solo la educación, estrategia a largo plazo, será capaz de moldear ese intelecto y sensibilidad que se requiere para la aprehensión de los grandes valores, incluidos los que solemos asociar con los derechos humanos. Estos también, nos gusten o no reconocerlo, dependen de una sensibilidad específica: la más elevada.

Por eso, creo que hay que insistir en la alfabetización digital, sí, pero no en el sentido en que suele proponerse. No se puede reintroducir las humanidades en las aulas en términos sumatorios: la IA, por un lado, y nuestra humanidad, por el otro. En esta equidistancia errónea, la humanización de la tecnología resulta incluso peligrosa. Véase dicha idea, por ejemplo, en un mantra que hoy cada vez más personas repiten: “*La IA ha venido para quedarse.*” Huxley –imagino– respondería: *puede que sí, o puede que no*. Pero también añadiría algo importante, una postura que comparto –la más optimista, la más esperanzadora–: Lo único que verdaderamente no pasará, lo único que no será dejado atrás, será el espíritu humano –irreductible. Desgraciadamente, como tantas veces en la historia, el coste de tener recordarlo esta idea –de vencer el error más siniestro– puede llegar a ser altísimo.



He elegido para terminar esta digresión la obra del *Prometeo encadenado*, también de Rubens –pintor, por cierto, de sólida formación humanista: del buen humanismo.

Rubens pinta el final del mito. Como castigo por su transgresión, Zeus, el rey de los dioses, ordenó que Prometeo fuera encadenado a una roca, donde un águila devoraba su hígado cada día, el cual volvía a regenerarse por la noche, en un ciclo de tormento eterno. Rubens anticipa aquí esa otra pintura goyesca, sobre los monstruos que es capaz de engendrar el sueño de la razón.

También quiero acompañar la pintura de unos pocos versos. Esta vez, compuestos por Miguel de Unamuno en 1910, para cantarle al buitre de Prometeo, buitre que ha hecho suyo porque, dice, siempre ha sido suyo (Unamuno 2013, 311).

¿Y después? ¿Cuándo cese el Pensamiento
de regir a los mundos?
¿Y después...?
—¡ay, ay, ay! ¡no tan recio! —
¡no tan recio, mi buitre!
mira que así me arrancas la conciencia;
aun dentro de tu oficio, ¡ten clemencia!

Para Unamuno, son los hiper-especialistas los principales responsables de que *cese el pensamiento de regir a los mundos*. No discrimina si de ciencias o de letras, todos, ¡todos! — bromea Unamuno— todos “caza-vocablos, barrunta-fechas y cuenta-gotas de toda laya” (Unamuno 2013, 211). Entonces, ¿quién está a salvo de este voraz buitre? Que no lo esté hoy

nadie, se consuela Unamuno, no es motivo para desesperar. Quizá lo puedan estar nuestros hijos.

¿Conmueve hoy el argumento de la prole? Algunos acusan a esta generación de despreocuparse por el futuro o, peor aún, de sacrificarlo –junto con sus propios hijos– en beneficio propio. No es un mal novedoso: ya aparece plasmado en el mito de Saturno, ese dios capaz de devorar a sus hijos para no perder el trono entre los dioses –hoy diríamos: entre los filósofos *influencers*, ávidos de *likes*. Lo realmente nuevo, quizá, pueda ser que la inteligencia artificial, actuando como una cámara de eco perfeccionada, se convierta en el adulator más eficaz, aquel que refuerza, valida y amplifica nuestras decisiones –incluso las más monstruosas. Un nuevo tipo de adulación en la corte del rey desnudo. Por eso también suele decirse que el silencio es el amigo del verdadero filósofo: nada que no sea realmente sólido resiste el silencio.

Bibliografía

Byung-Chul Han. *Rueda de prensa concedida a los medios antes de la recogida del premio Princesa de Asturias de Comunicación y Humanidades*. Publicado en el diario ABC el 21 de octubre de 2025. Consultado el 11 de octubre de 2025 en el siguiente link: <https://www.abc.es/cultura/byungchul-esperanza-colapso-sistema-pasar-pronto-20251021161916-nt.html>

de Botton A. *Interview by Alex O'Connor*. The Institute of Art and Ideas. Concedida el 16 de octubre de 2025. Consultado el 10 de octubre de 2025 en el siguiente link: <https://www.youtube.com/watch?v=2OUhzjtUXxE>

Goethe JW. *Poemas*. Madrid: Galaxia Gutenberg, 2002.

Hood L & Price N. *The age of scientific wellness*. Harvard University Press, 2023.

Huxley A. *La Isla*. Barcelona: Edhasa, 2003.

Huxley A. *Un mundo feliz*. Madrid: Cátedra, 2013.

Lipovetsky G. Entrevista concedida al Daniel Verdú para el diario *El País* el 21 de septiembre de 2025. Consultado el 11 de octubre de 2025 en el siguiente link: <https://elpais.com/ideas/2025-09-21/gilles-lipovetsky-si-quiere-vivir-mejor-enamorese-tome-prozac-no-busque-en-la-filosofia.html>

Panteli D et al. "Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions." *The Lancet Public Health*, 2025; 10(5): e428 - e432.

Platón. *Las leyes*. Madrid: Gredos, 1988.

Public Health. European Commission. "Artificial Intelligence in healthcare. 2025." Consultado el 10 de octubre de 2025 en el siguiente link: https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en#latest-updates-and-documents

Repullo Labrado JR. "La influencia de la tecnomedicina en la relación médico-paciente." En P. Martínez Jiménez et al. (Eds.), *Manual de la relación médico-paciente*. Foro de la Profesión Médica de España, 2024; 491-506.

Sagan C. *Interview by Charlie Rose*. Charlie Rose Show. PBS. Concedida el 27 de mayo de 1996. Consultado el 7 de octubre de 2025 en el siguiente link: <https://www.youtube.com/watch?v=iyFw8UF85A>.

Sarkar U, Bates D W. "Using Artificial Intelligence to Improve Primary Care for Patients and Clinicians." *JAMA Internal Medicine*, 2024; 184(4):343-344.

Unamuno M. *Niebla*. Madrid: Cátedra, 2013.

Unamuno. M. "Rosario de sonetos líricos." En *Poesía completa*. Volumen I. Madrid, Alianza, 1987.

Vigo A. "La labor del filósofo en la actualidad." Sesión impartida dentro del Coloquio permanente Instituto de Filosofía (USS) el 25 de septiembre de 2022. Consultado el 10 de octubre de 2025 en el siguiente link: <https://www.youtube.com/watch?v=xglmyO8LOUg>

WHO European Region. *Artificial intelligence is reshaping health systems: state of readiness across the WHO European Region*. Copenhagen: WHO Regional Office for Europe, 2025.

Yip W. "Improving primary healthcare with generative AI." *Nature Medicine*, 2024: 30, 2727–2728.

Žižek, Slavoj. *The Parallax View*. Cambridge, MA: MIT Press, 2006.

Žižek, Slavoj. *Welcome to the Desert of the Real*. London: Verso, 2002.

José Pedro García Sabater y Alfredo Marcos Martínez. ¿Se interesa la IA por los problemas del sur global?

José Pedro García Sabater

Catedrático de Organización de Empresas
Universidad Politècnica de València

Alfredo Marcos Martínez

Catedrático de Filosofía de la Ciencia
Uivesidad de Valladolid

[ACCEDE A LA GRABACIÓN](#)

Algunas consideraciones sobre la pregunta “¿Se interesa la IA por los problemas del Sur global?”

I

Más que intentar responder directamente a la pregunta inicial, lo que intentaré será analizarla, para después reformularla y esbozar, entonces, alguna línea de respuesta.

En primer lugar, nuestra pregunta inicial hace referencia a la IA, para plantear a continuación un problema ético-político relacionado con ella. Pues bien, desde mi punto de vista, los problemas de este tipo han de ser abordados solo después de haber pensado la naturaleza propia de la IA (lo que en filosofía llamaríamos su ontología).

Lo primero que hemos de preguntarnos es si la IA existe. No es una pregunta frívola. Muchos investigadores y desarrolladores de IA se hacen a diario esta cuestión. Existe algo a lo que llamamos IA, pero esta denominación no hace justicia a la verdad de la cosa así llamada. Como mínimo deberíamos anteponer siempre la palabra “sistema”. No existe la IA, existen diversos *sistemas de IA*. El matiz es importante. Hay filósofos de la tecnología que sostienen que todo sistema tecnológico está compuesto tanto por aparatos, como por personas. Es decir, un sistema de IA es un sistema compuesto por máquinas (software y hardware) y por personas (investigadores, desarrolladores, mantenedores, empresarios y accionistas, usuarios, legisladores, afectados...). Visto así, podemos afirmar que la parte inteligente del sistema la ponemos las personas, no los artefactos.

Con todo, creo que una denominación aún más ajustada a la realidad de la cosa es la de “sistemas de control delegado” (sistemas CoDe). De hecho, son sistemas en cuya parte mecánica delegamos las personas, total o parcialmente, el control de algunos procesos. El cambio de denominación está vinculado a un cambio radical de perspectiva. Pasamos de entender la IA como un conjunto de máquinas y algoritmos con implicaciones e impactos sociales, a entender los sistemas de control delegado como *sistemas sociales mediados por máquinas*. En un sistema social, lo importante son las relaciones entre personas. Las preguntas que de esta visión derivan tienen que ver sobre todo con la acción de delegar:

- ¿Quién delega?, ¿de quién es la mano que mece los algoritmos?
- ¿Está legitimado para hacerlo?
- ¿En qué sistemas delega?, ¿son apropiados y confiables?

- ¿Qué tipo de acciones o procesos son delegados?
- ¿De verdad son técnicamente/éticamente delegables?
- ¿Para cuánto tiempo queda delegado el control?
- ¿Es prudente obrar así?
- ¿Es reversible la delegación del control?
- ¿Al delegar el proceso, perdemos las habilidades correspondientes?
- ¿Qué sistemas de monitorización y evaluación están previstos?
- ¿Son estos adecuados y suficientes?
- ¿Qué riesgos corremos?, ¿quién corre más riesgos?
- ¿Qué daños se producen y a quién?
- ¿Es prudente asumirlos?
- ¿Qué ventajas obtenemos con la delegación del control?
- ¿Quién obtiene más ventajas?, ¿es justo que sea así?
- ¿Qué perdemos a cambio?, ¿quién lo pierde?, ¿es justo?
- ...

Ahora podemos reformular la pregunta inicial (recordemos: *¿Se interesa la IA por los problemas del Sur global?*). Con la nueva perspectiva, esta pregunta se convierte en esta otra: ¿Se interesan por el Sur global los propietarios, accionistas, investigadores, desarrolladores, reguladores... de los sistemas de control delegado? En términos más icónicos: ¿Le interesa el Sur global a Mark Zuckerberg, Elon Musk o Xi Jinping cuando desarrollan o aplican IA?, ¿le interesa a la Comisión Europea cuando redacta el *EU AI Act (Ley de inteligencia artificial de la Unión Europea)*?

El segundo término de la pregunta es el de “Sur global”. Es un término geopolítico que puede que sea útil en otros contextos, pero tal vez, en lo que se refiere a la reflexión sobre la IA, este término aporte más confusión que orientación. El Sur global incluye a una serie de países. Pero, ¿cuál es el rasgo común a todos ellos? Podría ser su situación geográfica, como indica la palabra “sur”, pero no es así. La mayor parte, con mucho, de la población de estos países vive en el hemisferio norte.

Podría referirse a la situación geográfica relativa. Por ejemplo, México está al sur de USA y Marruecos al sur de Europa. Sí, pero Corea del Norte está, obviamente, al norte de Corea del Sur. El primero de estos países pertenece al Sur Global y el segundo al Norte global. Australia pertenece al Norte global, como Nueva Zelanda, que están al sur de casi todos los países del llamado Sur global. Albania (Sur global) está al norte Grecia (Norte global), China (Sur global) está al norte de Taiwán (Norte global), Emiratos Árabes (Norte global) está al sur de Afganistán (Sur global)... y así podríamos encontrar otros muchos ejemplos. No es el sur absoluto, pero tampoco es el sur relativo lo que constituye un rasgo común a los países de referencia.

¿Será una división del mundo por criterios económicos de PIB o renta per cápita? Tampoco. Por ejemplo, Rumanía y Bulgaria (Norte global) tienen economías más pequeñas y débiles que China, India, Turquía, México, Brasil y otros muchos países del llamado Sur global.

Dicho de otro modo, hay países y zonas integradas en el llamado Sur global que son más ricas que otros países y zonas del llamado Norte global.

Apelemos, pues, al pasado colonial. Pero tampoco funciona como criterio. China, un país y un imperio desde tiempo inmemorial, está oficialmente incluida en el llamado Sur global, mientras que Corea del Sur, que ha sufrido diversas colonizaciones, algunas recientes, pertenece al llamado Norte global. También Taiwán, siempre bajo la amenaza de ocupación china, es Norte global. Japón, bajo supervisión de los Estados Unidos entre 1945 y 1952, se tiene por Norte global. Los países que fueron auténticas colonias de la URSS hace no mucho y que hoy están integrados en la UE son clasificados como Norte global, y así podríamos seguir.

Nos hemos quedado sin criterios para definir el llamado Sur global. Pero es que cuando hablamos de IA, la confusión es particularmente perturbadora. China es oficialmente parte del Sur global. Es el país, junto con la India, que más habitantes aporta al mismo. Es el líder del Sur global. Y es también la mayor potencia, junto con USA, en IA. Olvidar este dato cuando hablamos de IA es tanto como intentar obviar el elefante en la habitación (o en la cacharrería). Cuando constatamos, por ejemplo, que China posee minas de litio en Chile y en varios países de África, ¿qué hemos de decir?, ¿que la IA fabricada en el Sur global (China) se produce gracias a la explotación de las tierras y gentes del Sur global (Chile, Zimbabue, Malí, Sudáfrica, Zambia, Guinea, Angola y Nigeria...)? Si Huawei vende IA al gobierno de España, ¿qué diremos?, ¿que somos una colonia tecnopolítica del Sur global? Si los adolescentes europeos son adictos a digitales del Sur global...

Lo más probable es que tengamos que reformular de nuevo nuestra pregunta inicial para captar el genuino espíritu de la misma, prescindiendo ya del concepto de Sur global.

Vamos a intentarlo. Supongamos que, al hablar de Sur global, en realidad nos estamos refiriendo a las personas más vulnerables, vivan donde vivan, ya sea en las calles de San Francisco, asoladas por el fentanilo, en el barrio barcelonés de La Mina o en alguna periferia de Manila o de Kabul. Hablemos de personas vulnerables, ya sean una niña o un adolescente o un anciano de cualquier país del mundo. Pueden ser personas incluso de las que solemos ubicar en las clases medias.

La pregunta quedaría ahora así: ¿Se interesan por las personas más vulnerables los propietarios, accionistas, investigadores, desarrolladores, reguladores... de los sistemas de control delegado?

Mi tentación es responder con un simple NO. Después habría que entrar en matices, pero en líneas generales, el interés de quienes rigen la IA por los más vulnerables es realmente escaso. Baste como muestra una impactante entrevista realizada a Arturo Béjar, ex-directivo de Facebook, y emitida hace unos meses:

(https://www.lasexta.com/programas/salvados/arturo-bejar-exdirectivo-que-atrevio-denunciar-lado-oscuro-meta-son-conscientes-hacen-nada_20241013670c2951afcb530001e78d3a.html). En ella se denuncia el desinterés absoluto de Meta por los daños que Instagram está causando a niños y adolescentes de todo el mundo. O bien, véanse las alarmantes noticias sobre el sistema de control social implantado gracias a la IA por el gobierno comunista chino (<https://www.rtve.es/noticias/20230614/credito-social-nuevo-metodo-control-masivo-china/2447566.shtml>).

El desinterés de los poderosos de la IA por las personas vulnerables, rayano en lo criminal, es el mismo en Silicon Valley que en el Partido Comunista Chino. Las fuentes para el entrenamiento de modelos siguen siendo opacas, los sesgos siguen apareciendo, heredados de estas fuentes y heredables por los usuarios humanos, la evaluación de los modelos se hace ardua, las expectativas a menudo resultan interesadamente hinchadas, con el consiguiente riesgo de burbuja, la integración en el sistema productivo resulta dudosa (muchos proyectos se quedan en el llamado purgatorio de los proyectos piloto, solo un 5% llegan a fase de producción).

Algo más de deferencia hacia las personas vulnerables encontramos en la normativa de la UE, aunque, desde luego, resulte a la postre insuficiente, pues tiende a regular la IA como un producto y no como un poder. Además, a la UE se le puede afejar el hecho de que entre con ínfulas regulatorias en un territorio explorado y urbanizado previamente por otros (chinos y americanos). Incluso la regulación, que es necesaria, constituye un riesgo para los más vulnerables si no se hace bien. Al llegar necesariamente con retraso, ofrece a las *bigtech* el tiempo que necesitan para acumular un patrimonio que les permita hacer frente después con ventaja a las multas y demandas que puedan afectarles. Además, una regulación desmedida hace tan caro el desarrollo e implantación de sistemas de IA, que acaba resultando un mecanismo de protección a favor de las grandes tecnológicas y en contra de otras empresas más pequeñas y locales, que no pueden afrontar los gastos derivados de la propia regulación.

Por añadidura, la exigencia legal bienintencionada de *human-in-the-loop*, como garantía humanista, no siempre es realista y, además, traslada el problema a la educación de las personas. Por ejemplo, en los procesos de diagnóstico médico, la IA es de gran ayuda y coopera con el personal médico que la supervisa; pero, ¿qué pasará con la siguiente generación de médicos, formados ya con poca independencia de criterio respecto de la IA?, ¿tendrán el coraje de obrar incluso contra alguna de las sugerencias de diagnóstico o de tratamiento hechas por la IA cuando así lo vean conveniente?, ¿les respaldará la legislación? O bien, ¿cómo formar abogados jóvenes cuando los despachos entreguen las tareas que solían hacer los pasantes a sistemas de IA? Y así en cada dominio...

En el plano político las dudas preocupantes también se nos van acumulando. Por ejemplo, Emily Bender, reconocida lingüista de la Universidad de Washington, denuncia que “gracias a las enormes sumas invertidas, una minúscula camarilla de hombres tiene la capacidad de influir en lo que sucede a grandes franjas de la sociedad” (<https://www.xataka.com/robotica-e-ia/ia-no-que-bola-8-magica-pretensiones-emily-bender-linguista-academica>).

En una línea muy similar podemos citar también a Carissa Véliz, de la Universidad de Oxford. Según ella, la ética médica se desarrolló a partir de escándalos que hicieron muy tangible y muy obvio por qué la necesitábamos. Razones análogas hacen que la ética digital está surgiendo como una disciplina. Porque si todas estas empresas utilizaran prácticas éticas, no habría tanta necesidad de un cambio. La acusación de que si te preocupan las prácticas éticas es que estás en contra de la tecnología es pura retórica –dice– “porque estar en contra de la tecnología mal diseñada, de la tecnología que está dañando a la gente, no tiene nada que ver con estar en contra de la buena tecnología. Encontraríamos muy pocas personas contrarias a la tecnología por principio. Hay que diseñar mejor la tecnología para que respete los derechos humanos y para que sea una influencia positiva en la democracia y no negativa” (https://ethic.es/entrevistas/entrevista-carissa-veliz/?fbclid=IwY2xjawLOJZ1leHRuA2FlbQlxMQABHidBH5RrqIKYXolPkr6qGVtXcB8O7X3vBeQ29WKuTIUyY9McZY8HNNHgHRTaR_aem_M9ftCDp2aBNyDCGKJ41Mrg).

Entonces, ¿cómo plantarle cara a la amenaza que la inteligencia artificial supone para los más vulnerables sin caer en el neoludismo? Quizá la más efectiva de las estrategias para generar un uso sensato y humanitario de la IA, la mejor opción para producir una barrera de defensa que proteja a las personas vulnerables, resida, no principalmente en una regulación, sino en una educación en virtudes de todos los ciudadanos, en *tecnovirtudes* que estén a la altura de los actuales retos tecnológicos, incluidos los que provienen de la omnipresente digitalización. Entre estas tecnovirtudes señalaría con especial énfasis la serenidad o desasimiento y la prudencia. Y entre las prácticas que pueden ayudar a generarlas, señalaría la del silencio tecnológico. Pero, a cada día su afán. El desarrollo por extenso de estos dos conceptos, el de tecnovirtudes y el de silencio tecnológico ha de quedar para otra jornada.

Federico Buyolo García y Violeta Dávila. Consecuencias de la IA en el sector cultural

Federico Buyolo García

Director Cultural
Fundación Gregorio-Marañón

Violeta Dávila

Gestora cultural y cofundadora BeCooltural

[ACCEDE A LA GRABACIÓN](#)

El impacto de la Inteligencia Artificial en la cultura

El impacto de la inteligencia artificial en la cultura puede analizarse desde muchas perspectivas que iremos desarrollando. Pero antes conviene plantear una pregunta sencilla y, a estas alturas, nada inocente: ¿cómo pueden estar seguros de que las palabras que escucharon aquella mañana —o las que leerán ahora— no han sido generadas por una herramienta de inteligencia artificial generativa? Pronto será difícil encontrar un espacio de divulgación (¿podremos llamarlo de pensamiento?) que no surja de un proceso de creación asistido por inteligencia artificial.

Hecha la advertencia, y por transparencia, comparto que en la preparación de la presentación utilizamos una herramienta de IA -en concreto, ChatGPT- para esquematizar información, explorar dinámicas de conversación y, después, para curar y editar esta acta. Le pedimos expresamente que **no** escribiera por nosotros: queríamos conservar la voz y que el resultado fuese el de un proceso intelectual humano del que pudiéramos reivindicar la autoría.

El tema de la autoría está en construcción y aún tiene evidentes lagunas legislativas, pero quizá sea el debate más relevante. ¿Puede tener la inteligencia artificial derechos de autor? En España, la regla es clara: **solo un ser humano puede ser autor**. Lo dice el artículo 5 de la Ley de Propiedad Intelectual: la autoría corresponde a la persona física; una IA, por definición, no puede ostentar derechos de autor. Otra cosa es **quién explota** la obra: la legislación permite que una persona jurídica se beneficie patrimonialmente de una **obra colectiva** o de un **programa de ordenador**. En nuestro país las obras generadas íntegramente con IA no tienen un autor legal en sentido estricto. No obstante, cuando existe una **participación humana significativa**, puede plantearse la atribución de derechos.

Ahí aparece la zona gris: si el resultado “asistido por IA” incorpora una intervención humana significativa (definición de prompts, edición sustantiva, curaduría creativa). Según la -aún escasa- jurisprudencia los derechos de autor pueden corresponder al usuario si influye significativamente en el resultado.

A escala europea el panorama añade reglas sobre **cómo** se hace IA, no **quién** es autor. El Reglamento de IA exige respetar la propiedad intelectual existente y obliga a los proveedores de **modelos de uso general** a publicar **resúmenes de los datasets** y a **respetar las reservas de derechos**, pero evita entrar en el debate de la autoría de lo generado. Y la Directiva (UE) 2019/790 -el gran marco del entorno digital- habilita la **minería de datos** (clave para entrenar modelos) en ciertos supuestos, a la vez que arma a los titulares con un freno explícito: pueden **reservarse el derecho** y excluir sus obras del entrenamiento mediante señales como un *robots.txt* bien configurado.

Fuera de Europa, Estados Unidos adopta el mismo principio —la U.S. Copyright Office niega derechos de autor a obras generadas por IA—, mientras que países como China, Japón o el Reino Unido experimentan con fórmulas intermedias, sin llegar a reconocer a la IA como sujeto creativo. En síntesis: la IA no tiene personalidad jurídica ni capacidad creativa propia; no puede ser autora, aunque sí puede **ampliar** el ejercicio de la creación humana.

En Estados Unidos, *The New York Times* ha demandado a OpenAI y Microsoft por usar millones de artículos protegidos para entrenar modelos; en el Reino Unido, *Getty Images* acusa a Stability AI de utilizar su archivo -incluso con marcas de agua- para entrenar *Stable Diffusion*. Artistas gráficos han demandado a generadores por entrenar con obras sin permiso; editoriales y autores han puesto sobre la mesa el posible “Napster” de la literatura. En Francia, asociaciones de editores y autores han denunciado el entrenamiento con contenidos protegidos. La cuestión de fondo es si el entrenamiento constituye uso legítimo *-fair use-* o infracción, y si lo legítimo cambia cuando el material proviene de fuentes pirateadas. No cito aquí fallos en firme porque muchos procesos siguen en curso; sí subrayo esto: el ecosistema cultural se ha convertido en **laboratorio jurídico** de la IA.

Cuando hablamos de cultura conviene recordar lo que decía Raymond Williams (*Keywords*, 1976): “culture is one of the two or three most complicated words in the English language”. Cabría decirlo de todas las lenguas. Como recuerda Antonio Monegal: el debate sobre qué es “cultura” es inacabable y está siempre en construcción. Edward Burnett Tylor, en 1871, ofrecía la definición antropológica canónica: la cultura como “ese todo complejo que incluye conocimientos, creencias, arte, moral, costumbres y todas las demás capacidades y hábitos adquiridos por el ser humano como miembro de una sociedad”. Entre la concepción restringida (las artes bajo la protección de las musas) y la amplia (las tramas simbólicas que nos sostienen) oscilan casi todas las discusiones. Aquí asumo ambas: cultura como **sector** —editorial, audiovisual, música, artes escénicas, patrimonio, museos— y cultura como **ecosistema social** de símbolos, relatos y prácticas compartidas. La inteligencia artificial impacta a las dos.

¿Qué ha cambiado en la cultura? Mucho y muy deprisa. En el lado optimista, la IA ha **ampliado el acceso**: traducciones automáticas, subtitulado en tiempo real, lectura aumentada, audiodescripciones que permiten disfrutar de una obra a quien antes quedaba fuera. Ha **democratizado** la producción básica: hoy cualquiera, con una guía mínima, puede escribir un poema, una melodía, un cartel. Ha **reducido costes** y abierto herramientas nuevas de experimentación: edición más ágil, prototipado veloz, recombinación de estilos. En patrimonio, las posibilidades son enormes: digitalizar, reconstruir virtualmente, transcribir tradiciones orales, rescatar archivos fotográficos dañados. Cito un ejemplo cercano: el proyecto de la UPV para recuperar decenas de miles de fotografías afectadas por la DANA. También la cooperación ciudadana —*crowdsourcing*— encuentra en la IA un aliado para clasificar, describir y difundir,

y ya ha dado casos de éxito: el British Museum, con **MicroPasts**, movilizó a voluntarios para transcribir más de 30.000 fichas de objetos de la Edad del Bronce y realizar *photomasking* para modelado 3D; en España, la **Biblioteca Nacional** aplica IA en catalogación y análisis de fondos y, además, la IA fue clave para atribuir una obra desconocida a Lope de Vega tras transcribir automáticamente cientos de textos y detectar patrones estilísticos.

Pero la cara velada del espejo no es menos real. La primera tensión es la **autoría** -ya tratada-. La segunda, el **derecho a la cultura**. El artículo 27 de la Declaración Universal de Derechos Humanos y el artículo 15 del Pacto Internacional de Derechos Económicos, Sociales y Culturales reconocen el derecho a **participar** en la vida cultural y el derecho a **beneficiarse** del progreso científico. La IA empuja en ambas direcciones: puede ampliar acceso, participación, visibilidad y preservación; pero también concentrar el poder de mediación, homogeneizar lenguas y estéticas, erosionar la diversidad, desproteger los derechos de autor y aumentar las brechas de desigualdad. De nosotros depende orientar esa fuerza.

La tercera tensión es el proceso creativo. Se habla del “**artista aumentado**” (Philip Edwards, 2023): quien integra la inteligencia artificial como fuente de inspiración y como herramienta, no para reemplazar su creatividad, sino para **expandirla** y explorar conexiones, formas o ideas que quizá en solitario no imaginaría. Aun reconociendo esa potencia, el peligro no es -solo- que una máquina “cree mejor” en ciertos parámetros, sino que **nos deshabitúe del proceso**. Si la creación se reduce a una fórmula de *prompt + filtro + curaduría*, corremos el riesgo de devaluar la búsqueda, el error, la deriva que encuentra un tono propio y, poco a poco, **delegar la imaginación**. Esa dependencia cognitiva merma el potencial creativo humano. ¿Y si hay un apagón, qué hacemos? En definitiva, la IA devuelve **respuestas**; la creatividad habita en las **preguntas**. (Quizá Sócrates hoy se sentiría en su salsa conversando con ChatGPT).

Hay riesgos **económicos y laborales** visibles: desplazamiento de tareas, precarización de profesiones ya frágiles (traductores, correctores, ilustradores, músicos), concentración de valor en plataformas. Circulan estimaciones -en música, pérdidas potenciales; en audiovisual, presión a la baja en guion y posproducción- que, sin ser oráculos, marcan una tendencia preocupante para el sector. También hay **brecha de competencias**: artistas con oficio a los que va a resultar más difícil adaptarse a este nuevo escenario. Y está la **homogeneización**: los algoritmos recomiendan lo que “funciona”, y lo que “funciona” acaba formateando lo que se produce. Riesgo de **cultura sin experiencia**, cultura de superficie, cultura sin fricción.

Añadamos un capítulo que preocupa -con razón-: **deepfakes y desinformación visual**. La IA ya permite falsificar con verosimilitud rostros y voces. En España, el artículo 18 de la Constitución protege el honor, la intimidad y la propia imagen; la LO 1/1982 refuerza la prohibición de alterar o difundir imágenes sin consentimiento; la LOPDGDD protege datos personales, y la LSSI establece responsabilidades de los prestadores de servicios digitales ante

contenidos ilícitos. Sirve, pero no basta. El **anteproyecto de Ley de IA** (marzo de 2025) apunta a clasificar como de **alto riesgo** las tecnologías que generen contenidos manipulativos, imponiendo obligaciones de **transparencia, trazabilidad y actuación diligente** a plataformas. Es un paso en la dirección correcta, con dos retos abiertos: la **jurisdicción global** (el contenido puede producirse en cualquier país) y el grado de **responsabilidad** que exigiremos a plataformas que operan a escala planetaria.

A pesar de todo, no comparto el fatalismo. La IA, bien encauzada, se vuelve **aliada**. En la accesibilidad, permite disfrutar de aquello que antes quedaba vedado. En la preservación, salva archivos, reconstruye pigmentos perdidos, da segunda vida a voces y músicas. En la mediación, puede acercar obras minoritarias a públicos que nunca las habrían encontrado. En la **revitalización lingüística**, imaginemos modelos que traduzcan y generen en lenguas minorizadas con la misma dignidad que en inglés: no para sustituir a nadie, sino para sostener una comunidad de voces.

Detengámonos en el **Sur Global**, porque el congreso invita a mirar desde ahí. La infraestructura -datos, cómputo, nubes- está concentrada en el Norte y en China. Los modelos se entrenan mayoritariamente en inglés, con corpus que reproducen sesgos culturales y de género; lenguas indígenas o locales aparecen poco o nada. Esto limita la calidad de las respuestas, pero sobre todo **limita la imaginación**: lo que no está en el corpus, no existe para la máquina; lo que no existe para la máquina, circula menos; lo que circula menos, se debilita. La **dependencia tecnológica** -cuando para crear, distribuir o cobrar dependes de un puñado de plataformas- se vuelve **dependencia cultural**. Y no solo: hay riesgos de **vigilancia** algorítmica en contextos con democracias frágiles; hay **desinformación barata** en procesos electorales; hay **fuga de talento** hacia centros que pueden pagar el cómputo que en otros lugares es inaccesible. La respuesta no puede ser el repliegue, sino la **gobernanza**: cooperación internacional, licencias justas de entrenamiento, fondos para infraestructura pública, soporte a proyectos de IA **en y para** el Sur, incorporación de comunidades locales a la toma de decisiones.

Todo esto obliga a un **marco de principios**. En Mondiacult 2025 se enumeran con claridad: la IA debe **servir a la cultura, no sustituirla**; debe **amplificar** la creatividad humana; debe **garantizar trazabilidad y proteger a los creadores** -también frente al uso masivo de sus obras para entrenar modelos-; debe **salvaguardar** la diversidad cultural y lingüística; su **gobernanza** ha de ser **participativa**; y debemos invertir en **educación, investigación e inclusión** para que nadie quede atrás. Sumemos criterios prácticos: **etiquetado** claro (saber cuándo un contenido está generado o asistido por IA), **registro** de procesos (para repartir valor donde corresponda), **licencias** de entrenamiento que remuneren a titulares cuando proceda, **evaluaciones de impacto** cultural y de derechos en instituciones públicas.

No escondo mis temores, porque nombrarlos es la mejor forma de ponerles límites. Me alarma que, por comodidad o fascinación, cedamos demasiado terreno. Que, en nombre de la eficiencia, recortemos lo más valioso que tiene un oficio: su **proceso**. Que deleguemos la imaginación y nos contentemos con curar resultados. Que los equipos se acostumbren a trabajar “para el algoritmo” y no para un lector, un espectador, una comunidad. Que confundamos **abundancia** con **riqueza**: producir más no es crear mejor; y un océano de contenidos no garantiza una cultura más viva. También me preocupa la **economía del sector**: si el valor añadido se acumula en la capa tecnológica y no en la creativa o la de mediación, ¿cómo sostendremos la profesión? ¿qué pasará con quienes hoy viven -apenas- de corregir, traducir, ilustrar, programar? No es catastrofismo; es pedir **reglas** que repartan de manera justa.

Hay también **motivos para la esperanza**, y conviene sostenerlos. Hay museos que usan IA para acercar obras a personas con discapacidades sensoriales; archivos que, con herramientas abiertas, rescatan memorias de barrios enteros; comunidades que usan modelos para documentar y transmitir su patrimonio oral. Y he visto —en el día a día— escritoras, músicos, editoras que aprenden a usar estas herramientas **sin perder su voz**, como quien incorpora un instrumento más a la orquesta. Esa es, quizá, la clave: no deslizar la responsabilidad hacia la máquina, sino **dirigirla**. Poner a la IA en su lugar: herramienta poderosa, no oráculo; aliada exigente, no sustituta.

Vuelvo al inicio. La pregunta por **quién habla** —si una persona o una máquina— no es un recurso retórico; es un punto de partida ético. Debemos reforzar la **confianza**, rendir cuentas y reivindicar la **transparencia**. Importa saber quién hay detrás de un texto, una canción, una obra visual. Qué puede contarnos ChatGPT sobre la muerte de un padre; cómo puede emocionarnos, en cambio, Gueorgui Gospodínov en *El jardinero y la muerte*.

Mientras eso exista, la cultura seguirá siendo nuestro espacio común: el lugar donde nos contamos quiénes somos y quiénes queremos ser. Y la IA, con sus luces y sus sombras, tendrá cabida. Pero si renunciamos a esa **experiencia** y aceptamos un sucedáneo de superficie, entonces no será la máquina quien nos haya vencido, sino nuestra pereza.

La cultura es, antes que nada, un ejercicio de **presencia, se crea en lo físico, en la materialidad**. Y como dice el filósofo Rob Riemen es siempre el resultado de una conciencia moral. Que las máquinas hagan su parte -la que les toque-, pero que no nos roben la nuestra: imaginar, emocionar, expresar, agitar. Si el arte nace de la experiencia humana, defendamos esa fuente. Ante la pregunta ¿hay cultura si no es humana? la respuesta es: **sin experiencia humana no hay cultura posible; y sin cultura, no hay futuro**.

Bibliografía

Tylor, E. B. (1871). *Primitive Culture*.

Williams, R. (1976). *Keywords: A Vocabulary of Culture and Society*.

Monegal, A. (2022). *Como el aire que respiramos*. Ed. Acantilado.

Documento Final MONDIACULT Compromiso de los Ministros y Ministras de Cultura
MONDIACULT 2025

Report of the Independent Expert Group on 2025 Artificial Intelligence and Culture. Unesco.

Concepción Campos Acuña e Idoia Salazar García. Innovación social en la IA: ¿un espacio entre la regulación y la competitividad?

Concepción Campos Acuña

Presidenta de la Asociación MPS
Profesora de Derecho Administrativo (URV)
Experta en Gestión Pública

Idoia Salazar García

Presidenta del Observatorio del Impacto Social y Ético de la IA, OdiselA
Profesora de Periodismo. Universidad San Pablo CEU

[ACCEDE A LA GRABACIÓN](#)

**Aída Ponce del Castillo y Adrián Todolí Signes. Reemplazo y deslocalización
del recurso humano. Retos para el futuro del trabajo en la era del
capitalismo digital**

Aída Ponce del Castillo

Investigadora. European Trade Union Institute

Adrián Todolí Signes

Profesor de Derecho del Trabajo y de la Seguridad Social
Universitat de València

[ACCEDE A LA GRABACIÓN](#)

Lorenzo Cotino Hueso. ¿Es posible una gobernanza mundial sobre la IA?

Lorenzo Cotino Hueso

Presidente de la Agencia Española de Protección de Datos

[ACCEDE A LA GRABACIÓN](#)

Comunicaciones

COMUNICACIONES PRESENCIALES

- *Una experiencia metodológica en la enseñanza de la IA y los derechos humanos*
Guadalupe Bohorques Marchori. Universidad Europea de Valencia
- *La Inteligencia Artificial generativa y el Artículo 27 de la Declaración Universal de los Derechos Humanos*
Luis Demano. En representación de APIV (Associació de Professionals de la Il·lustració Valenciana)
- *Inteligencia artificial, accesibilidad y diseño inclusivo en el metaverso: hacia una justicia tecnológica en la era 5IR*
Dra. Chele Esteve Sendra, Universitat Politècnica de València. **Ángela Muñoz Iglesias,** Universitat Politècnica de València. Dr. **Rafel Arnal Rodrigo,** Universidad de Zaragoza.
- *IA y la libertad religiosa*
Antonio Quirós-Fons. Universidad Europea de Valencia.
- *Inteligencia artificial y acoso escolar: responsabilidad civil de los centros por hechos de alumnos bajo su guarda*
Noelia Liduina Rebón Rodríguez
- *From Predicting the Past to Designing the Future: A Digital Twin Vision for Human Services*
Kaj Wik Siebert. Independent Consultant, Data & AI for Social Impact.

Guadalupe Bohorques Marchori. Una experiencia metodológica en la enseñanza de la IA y los derechos humanos

Guadalupe Bohorques Marchori
Universidad Europea de Valencia

[ACCEDE A LA GRABACIÓN](#)

Resumen

Este capítulo busca abordar los dilemas éticos y de derechos humanos que plantea la implementación de IA, desde una ética integral. En primer lugar, se analiza la evolución y aplicación de la IA en el ámbito académico, explorando sus avances, oportunidades y limitaciones. Después, se describe detalladamente una actividad práctica realizada en el aula que invita a reflexionar sobre los dilemas éticos asociados al uso de la IA, con un enfoque particular en los derechos humanos. A continuación, se exponen los resultados de esta experiencia educativa, destacando sus beneficios, limitaciones y posibles mejoras. Finalmente, se ofrecen conclusiones que invitan a reflexionar sobre la importancia de encontrar un equilibrio entre la innovación tecnológica y la responsabilidad ética, proponiendo caminos para una integración responsable y ética de la IA en la educación.

INTRODUCCIÓN

La ética ha sido un pilar fundamental en la construcción de sociedades justas, equitativas y responsables. Desde sus orígenes filosóficos, la ética ha buscado reflexionar sobre la conducta humana, delimitando lo correcto e incorrecto y estableciendo principios que guían nuestras decisiones y acciones (Floridi, 2013) “un arte de vivir”. Esta disciplina se ha convertido en un componente esencial para analizar y regular el impacto de los avances científicos y tecnológicos, especialmente en un mundo donde el progreso se desarrolla a un ritmo acelerado. En este contexto, la ética aplicada, entendida como una rama de la ética contemporánea que se aplica a cuestiones concretas de la vida y a la actividad profesional (Bohorques Marchori, 2014), adquiere un papel determinante al abordar los desafíos que surgen con el desarrollo de nuevas tecnologías.

En sectores como la medicina, la IA ha permitido diagnósticos más precisos y tratamientos personalizados (Jiang et al., 2017); en la economía, ha optimizado la gestión de recursos y predicciones de mercado (Erik & Andrew, 2017); y en la educación, ha prometido revolucionar los métodos de enseñanza y aprendizaje (Gómez, 2023). No obstante, estos avances han despertado importantes cuestionamientos éticos y sociales, especialmente en ámbitos donde la creatividad, la sensibilidad cultural y el juicio humano son esenciales (Topol, 2019). La posibilidad de que sistemas automatizados influyan en decisiones que afectan a individuos y comunidades genera incertidumbre y plantea dilemas éticos complejos (IEEE, 2019).

Particularmente en el campo educativo, la incorporación de la IA ha abierto debates sobre su potencial para enriquecer los procesos de enseñanza-aprendizaje (Aparicio-Gómez & Gallego, 2024; Holmes, 2019), así como sobre sus riesgos y limitaciones. La educación no solo consiste en la transmisión de conocimientos, sino también en la formación de ciudadanos críticos, responsables y éticamente comprometidos. La integración de tecnologías de IA en las aulas implica cuestionarse sobre la autonomía del aprendizaje, la equidad en el acceso a recursos digitales, la privacidad de los datos personales y la posible deshumanización de las relaciones educativas. La aplicación de la IA en contextos educativos puede generar beneficios significativos, como la personalización del aprendizaje (Luckin et al., 2016), pero también podría reforzar desigualdades y disminuir la interacción humana (Selwyn, 2019), elementos esenciales para el desarrollo integral del estudiante.

Este capítulo busca abordar estos dilemas desde una perspectiva ética integral. En primer lugar, se analiza la evolución y aplicación de la IA en el ámbito académico, explorando sus avances, oportunidades y limitaciones. Después, se describe detalladamente una actividad práctica realizada en el aula que invita a reflexionar sobre los dilemas éticos asociados al uso de la IA, con un enfoque particular en los derechos humanos, que integra el conocimiento y límites de aplicación de la reciente legislación europea en esta materia. A continuación, se exponen los resultados de esta experiencia educativa, destacando sus beneficios, limitaciones y posibles

mejoras. Finalmente, se ofrecen conclusiones que invitan a reflexionar sobre la importancia de encontrar un equilibrio entre la innovación tecnológica y la responsabilidad ética, proponiendo caminos para una integración responsable y ética de la IA en la educación.

Diferencia entre problema y dilema

Para comprender de manera profunda los desafíos éticos que plantea la inteligencia artificial (IA), resulta fundamental establecer una clara distinción entre los conceptos de problema y dilema. Esta diferenciación no solo es relevante en términos teóricos, sino que también es crucial para abordar de forma adecuada las complejidades que emergen con la implementación de tecnologías avanzadas en distintos ámbitos, en este caso en la educación (Stahl, 2021).

Un problema es, en esencia, una situación que puede resolverse mediante la aplicación de soluciones técnicas, lógicas o metodológicas. Los problemas suelen tener una o varias respuestas correctas y pueden ser abordados a través de procedimientos claros y definidos. Por ejemplo, si un software educativo presenta errores en la personalización de los contenidos para los estudiantes, este inconveniente puede resolverse ajustando los algoritmos, corrigiendo líneas de código o mejorando el diseño del programa. En estos casos, la solución es tangible y concreta, y su aplicación generalmente no involucra un conflicto de valores o principios éticos. La solución de problemas se centra en la eficacia, la precisión técnica y la optimización de procesos.

Por el contrario, un dilema implica tomar decisiones complejas donde ninguna de las opciones disponibles es completamente correcta o incorrecta. Los dilemas suelen involucrar conflictos entre valores, principios éticos o intereses contrapuestos, lo que hace que la decisión adecuada dependa de una profunda reflexión. En este contexto, cada posible acción tiene consecuencias éticas, sociales o personales que deben ser cuidadosamente sopesadas (Friedman & Hendry, 2019). Los dilemas no pueden resolverse aplicando reglas o procedimientos estándar, ya que requieren una evaluación crítica y un juicio ético. Un ejemplo clásico de dilema es decidir entre priorizar la privacidad de los usuarios o la seguridad pública al implementar sistemas de vigilancia masiva. Ambas decisiones tienen beneficios y riesgos, pero ninguna solución es absoluta (UNESCO, 2021).

En el ámbito ético, esta distinción entre problema y dilema cobra especial relevancia (Schiff et al., 2021). Los problemas éticos, al igual que los problemas técnicos, pueden resolverse siguiendo normas, leyes o códigos de conducta previamente establecidos (Jobin et al., 2019). Sin embargo, los dilemas éticos exigen un análisis más profundo y una deliberación cuidadosa, ya que involucran cuestiones fundamentales relacionadas con la dignidad humana, los derechos individuales y el bienestar social. Los dilemas éticos no tienen soluciones claras ni objetivas, sino que requieren ponderar cuidadosamente los principios y valores en conflicto para tomar una decisión responsable.

Aplicando esta distinción al uso de la IA en la educación, se puede afirmar que este contexto plantea un dilema ético complejo más que un simple problema técnico (Seldon et al., 2020). La cuestión no radica únicamente en mejorar los algoritmos de los sistemas educativos o en optimizar sus funcionalidades. El verdadero dilema surge al cuestionar si es moralmente aceptable permitir que la IA asuma un rol activo en la enseñanza, posiblemente desplazando la interacción humana, que es esencial para el desarrollo integral de los estudiantes. Este dilema implica considerar hasta qué punto la automatización y la inteligencia artificial pueden sustituir aspectos fundamentales de la educación, como la empatía, la creatividad, el juicio crítico y la transmisión de valores humanos (Al-Zahrani & Alasmari, 2024).

La eficiencia tecnológica que ofrece la IA es innegable. Los sistemas de IA pueden personalizar el aprendizaje, adaptarse a las necesidades individuales de los estudiantes y ofrecer recursos educativos de forma inmediata. Sin embargo, este beneficio debe evaluarse frente a la posible pérdida de calidad educativa y la deshumanización del proceso de aprendizaje (Cukurova, 2024). La interacción con docentes, el trabajo en equipo y el desarrollo de habilidades sociales son elementos fundamentales que difícilmente pueden ser reemplazados por sistemas automatizados. Además, existe el riesgo de que los sistemas de IA perpetúen sesgos existentes, discriminen a ciertos grupos o limiten la diversidad de perspectivas en el proceso educativo (Morley et al., 2020).

Reconocer este dilema ético es crucial para abordar de manera responsable la implementación de la IA en la educación. Esto requiere establecer límites claros en el uso de estas tecnologías, garantizando que actúen como herramientas complementarias y no sustituyan el papel insustituible de los educadores. También es necesario desarrollar marcos regulatorios adecuados que protejan la integridad del proceso educativo, salvaguarden los derechos de los estudiantes y promuevan la equidad en el acceso a recursos tecnológicos (Miao et al., 2023).

Además, resulta fundamental fomentar un diálogo constructivo entre todos los actores implicados: desarrolladores de tecnología, profesionales de la educación, expertos en ética, organismos reguladores, alumnos y familias. Este diálogo debe centrarse en encontrar un equilibrio entre la innovación tecnológica y el respeto por los valores humanos fundamentales. Solo mediante una colaboración activa y transparente será posible diseñar soluciones que integren la IA de manera ética y responsable, beneficiando a la educación sin comprometer su esencia humana (Al-Zahrani & Alasmari, 2024).

En definitiva, el desafío no consiste en rechazar la tecnología, sino en incorporarla de forma crítica y reflexiva, asegurando que su uso contribuya al desarrollo integral de los estudiantes y respete los principios éticos que sustentan el proceso educativo. Este equilibrio permitirá aprovechar los beneficios de la IA sin poner en riesgo los valores esenciales de la enseñanza y el aprendizaje.

OBJETIVOS

El presente capítulo tiene como objetivo fundamental distinguir entre problemas y dilemas éticos, subrayando la importancia de comprender que los primeros suelen resolverse a través de normas, leyes o códigos preestablecidos, mientras que los segundos exigen un análisis más profundo y una deliberación que involucre valores, principios morales y posibles consecuencias a largo plazo. En este sentido, la relevancia de los valores humanos en la adopción de la inteligencia artificial (IA) resulta innegable: la dignidad humana, los derechos individuales y la búsqueda del bienestar social se erigen como pilares que deben guiar cualquier decisión relacionada con la implementación de tecnologías en la educación. A lo largo de este trabajo, se enfatiza la necesidad de evaluar con detenimiento el impacto de la IA en la enseñanza y el aprendizaje, reconociendo las ventajas que ofrece al tiempo que se alerta sobre las potenciales consecuencias negativas, entre ellas la deshumanización del proceso formativo, la reducción de la interacción social y la posible perpetuación de sesgos o discriminaciones.

En esta línea, y con el fin de hacer frente a los desafíos que plantea la IA, se considera una nueva metodología: el Aprendizaje Basado en Dilemas (ABD), concebido como una variante innovadora del Aprendizaje Basado en Problemas (ABP). Mientras que el ABP tiende a enfocarse en la resolución de situaciones prácticas mediante metodologías o soluciones técnicas, el ABD desplaza el eje hacia la reflexión profunda de cuestiones éticas, valorativas y humanas que no admiten respuestas únicas. Esta propuesta, de creación propia, implica que el alumnado se enfrente a escenarios donde las decisiones conllevan ponderar principios morales, derechos fundamentales y efectos sociales a largo plazo, lo que resulta especialmente pertinente en el ámbito de la IA.

Asimismo, dentro de los objetivos de esta iniciativa formativa se incluye dar a conocer la futura legislación europea sobre Inteligencia Artificial, la primera en el mundo que entrará en vigor en 2026. Este marco legal afectará de manera significativa el desarrollo y la aplicación de la IA en diversos sectores, incluido el educativo. Por ello, uno de los propósitos centrales de la actividad es el alumno comprenda las implicaciones de esta normativa, adoptando una postura crítica y responsable ante los retos y oportunidades que traerá consigo.

Para hacer frente a los desafíos descritos, se subraya la conveniencia de establecer directrices éticas y regulatorias claras que garanticen la protección de la integridad del proceso educativo, salvaguarden los derechos de los estudiantes y aseguren la equidad en el acceso a las tecnologías de IA. De igual modo, se enfatiza la importancia de fomentar un diálogo constructivo y multidisciplinario que involucre a desarrolladores de tecnología, docentes, expertos en ética, organismos reguladores, alumnos y familias, con el fin de diseñar soluciones equilibradas respetando los valores esenciales de la educación.

Esta iniciativa subraya la importancia de incorporar la IA en la educación de manera crítica y reflexiva, sin menoscabar la empatía ni la equidad. Mediante el ABD, se propone formar estudiantes conscientes de los retos éticos de la tecnología, dotándolos de las habilidades necesarias para decidir con responsabilidad y valores sólidos.

METODOLOGÍA

La actividad titulada "Inteligencia artificial y derechos humanos: dilemas éticos" fue diseñada con el propósito de sensibilizar a los estudiantes sobre los desafíos éticos que plantea el uso de la inteligencia artificial (IA) en distintos ámbitos de la sociedad. Esta experiencia educativa no solo buscó transmitir conocimientos teóricos sobre la IA, sino también desarrollar habilidades de pensamiento crítico, análisis ético y reflexión sobre los derechos humanos. La actividad se estructuró en cinco fases, cada una enfocada en explorar diferentes dimensiones éticas y sociales de la IA.

Fase 1: Análisis de sesgos en los algoritmos

La primera fase de la actividad consistió en explorar el funcionamiento de los algoritmos y su impacto en la percepción de la realidad. Se les pidió que buscaran a las personas más influyentes de la historia y los mejores artistas. Este hallazgo llevó a un debate crítico sobre la supuesta neutralidad de la IA, cuestionando hasta qué punto los algoritmos reflejan o perpetúan prejuicios sociales. Los estudiantes analizaron cómo los datos de entrenamiento de los algoritmos pueden estar sesgados y cómo esto impacta en la diversidad y representación de distintas culturas y géneros. La discusión también abordó la responsabilidad de las empresas tecnológicas en la creación de sistemas más inclusivos y diversos.

Fase 2: Debate sobre autoría y creatividad de la IA

En la segunda fase, se presentó a los estudiantes el poema "The Stolen Time" y la obra visual "Théâtre D'opéra Spatial", ambas creadas por sistemas de inteligencia artificial. La finalidad de este ejercicio era abrir un debate sobre la autoría y la creatividad en obras generadas por IA. Los estudiantes reflexionaron sobre si una máquina puede ser considerada autora de una obra artística y cuáles serían las implicaciones legales y éticas de reconocer derechos de autor a un algoritmo.

El debate se enriqueció al discutir cómo la creatividad humana difiere de la producción automatizada. Se plantearon preguntas sobre la originalidad, la inspiración y la intención detrás de las obras, conceptos tradicionalmente asociados a la creatividad humana. Los estudiantes también analizaron el impacto que podría tener la IA en las industrias creativas y culturales, especialmente en términos de derechos de autor y reconocimiento de la labor humana.

Fase 3: Análisis de casos éticos reales

La tercera fase se enfocó en el análisis de casos reales donde la inteligencia artificial (IA) ha generado controversias éticas. Los estudiantes examinaron situaciones como la denegación de beneficios sociales debido a errores en sistemas automatizados, afectando a personas vulnerables; detenciones injustas causadas por fallos en el reconocimiento facial, evidenciando sesgos raciales y de género; la clonación de voces de artistas para crear canciones sin su consentimiento, planteando dilemas sobre derechos de autor y uso de imagen; y los conflictos de responsabilidad legal en accidentes con vehículos autónomos. Estos casos permitieron reflexionar sobre los derechos humanos vulnerados y discutir soluciones, subrayando la importancia de establecer regulaciones claras y garantizar la supervisión humana en el desarrollo y uso de la IA.

Fase 4: Reflexión sobre el impacto de la IA en la educación

La cuarta fase invitó a los estudiantes a reflexionar sobre el impacto de las tecnologías de IA en el ámbito educativo. Se discutió el uso de herramientas como ChatGPT y otros sistemas de IA aplicados al aprendizaje. Los estudiantes analizaron cómo estas tecnologías pueden mejorar la personalización del aprendizaje, facilitar el acceso a la información y optimizar los procesos educativos.

Sin embargo, también se debatieron los riesgos asociados, como la posible dependencia tecnológica, la desinformación y la pérdida de habilidades de pensamiento crítico. Los estudiantes reflexionaron sobre la necesidad de implementar estas tecnologías de forma responsable, asegurando que complementen, pero no sustituyan, la interacción humana y el rol del docente.

Fase 5: Explicación de la legislación europea

En esta fase, se explicó a los estudiantes el marco legal vigente en Europa respecto al uso de la inteligencia artificial. Se presentó el contenido del Reglamento Europeo de Inteligencia Artificial (Parlamento Europeo y del Consejo, 2024), que establece directrices claras sobre lo que está permitido y prohibido en el uso de estas tecnologías. Los estudiantes analizaron cómo esta legislación busca equilibrar el desarrollo tecnológico con la protección de los derechos fundamentales, garantizando la seguridad, el honor y la intimidad de las personas, la transparencia y la supervisión humana en sistemas de IA.

Fase 6: Propuestas de regulación y uso ético de la IA

En la última fase, los estudiantes se organizaron en grupos para desarrollar propuestas que abordaran los dilemas éticos analizados durante la actividad. Diseñaron soluciones enfocadas en regular el uso de la IA, priorizando la transparencia, la equidad y la responsabilidad. Algunas

propuestas incluyeron la implementación de auditorías éticas en el desarrollo de algoritmos, la creación de leyes que protejan los derechos de autor frente a la creación automatizada, y la exigencia de supervisión humana en procesos automatizados críticos.

Esta última fase permitió a los estudiantes aplicar sus conocimientos y reflexionar activamente sobre cómo pueden contribuir a la creación de un entorno tecnológico más ético y responsable.

RESULTADOS

La implementación de la actividad "Inteligencia artificial y derechos humanos: dilemas éticos" tuvo un impacto profundo y significativo en el aprendizaje y desarrollo crítico de los estudiantes, transformando su percepción sobre el papel de la inteligencia artificial en la sociedad. A lo largo de la actividad, se observaron avances sustanciales en su capacidad para analizar, debatir y reflexionar sobre los complejos dilemas éticos que surgen del uso de tecnologías emergentes. Este proceso de aprendizaje no solo amplió sus conocimientos teóricos, sino que también fortaleció habilidades esenciales como el pensamiento crítico, la argumentación lógica y la capacidad de proponer soluciones innovadoras a problemas reales.

Desde el inicio de la actividad, los estudiantes demostraron una comprensión más profunda sobre cómo los algoritmos influyen en la percepción de la realidad. Al realizar búsquedas en Google sobre "Las personas más influyentes de la historia" y "Los mejores artistas del siglo XX", identificaron patrones de sesgos que favorecían a figuras masculinas y occidentales, evidenciando una falta de diversidad y representación en los resultados. Este hallazgo desencadenó reflexiones críticas sobre la supuesta neutralidad de la inteligencia artificial y llevó a cuestionar cómo los datos de entrenamiento pueden perpetuar estereotipos y desigualdades estructurales. La discusión permitió a los estudiantes comprender la importancia de desarrollar algoritmos más inclusivos y éticos, así como la necesidad de responsabilizar a las empresas tecnológicas por las decisiones que toman al diseñar y aplicar estos sistemas.

El análisis de obras artísticas generadas por IA, como el poema *The Stolen Time* y la imagen *Théâtre D'opéra Spatial*, enriqueció aún más la experiencia educativa. La presentación de estas obras suscitó debates profundos sobre el concepto de autoría, creatividad y derechos de propiedad intelectual. Los estudiantes exploraron si un algoritmo, que carece de conciencia e intención, puede ser considerado autor de una creación artística. Este cuestionamiento llevó a analizar el impacto que estas tecnologías pueden tener en el reconocimiento y valoración de la creatividad humana. También se discutió la posibilidad de que el uso de IA en la creación artística pueda desplazar a los artistas humanos y transformar industrias creativas enteras, afectando la diversidad cultural y la autenticidad de las expresiones artísticas. Esta reflexión incentivó a los estudiantes a valorar la importancia de proteger los derechos de autor y la necesidad de establecer límites claros en la participación de la IA en procesos creativos.

La exposición a casos reales donde la IA ha tenido consecuencias negativas sobre los derechos humanos fue un componente esencial para sensibilizar a los estudiantes. Ejemplos como los errores en sistemas de reconocimiento facial que resultaron en detenciones injustas, o la clonación de voces de artistas para crear música sin consentimiento, permitieron visibilizar los riesgos asociados con el uso indiscriminado de la IA. Estos casos llevaron a un análisis crítico sobre la responsabilidad de las empresas tecnológicas, los vacíos legales existentes y la urgencia de establecer marcos regulatorios que protejan los derechos fundamentales.

Finalmente, la fase de creación de propuestas para mitigar los riesgos asociados al uso de la IA potenció la participación activa y el trabajo colaborativo. Los estudiantes, organizados en grupos, diseñaron soluciones innovadoras orientadas a promover un uso más ético y transparente de la IA. Entre sus propuestas destacaron la necesidad de establecer marcos legales sólidos que regulen el desarrollo y uso de algoritmos, implementar auditorías éticas periódicas para supervisar el funcionamiento de los sistemas de IA y garantizar la supervisión humana en procesos automatizados críticos. También sugirieron la creación de campañas educativas para sensibilizar a la sociedad sobre los riesgos de la IA y la importancia de la ética en su desarrollo.

Esta experiencia permitió a los estudiantes no solo comprender la dimensión técnica de la IA, sino también asumir una postura crítica y responsable frente a su uso. La actividad les brindó herramientas para identificar riesgos éticos, proponer soluciones y valorar la importancia de un desarrollo tecnológico que respete los derechos humanos. En concreto, la actividad promovió un aprendizaje integral que prepara a los estudiantes para enfrentar los desafíos del mundo digital desde una perspectiva ética y responsable.

Beneficios de la actividad

La actividad desarrollada ofreció múltiples beneficios educativos y formativos que impactaron positivamente en el aprendizaje y el desarrollo personal de los estudiantes. Uno de los principales logros fue el fomento del pensamiento crítico. Los participantes aprendieron a cuestionar la información presentada por los sistemas de inteligencia artificial, identificando sesgos en los algoritmos y analizando sus implicaciones sociales y culturales. Esta capacidad crítica les permitió evaluar de manera más profunda el impacto de las tecnologías emergentes en distintos contextos sociales. Al reflexionar sobre cómo los algoritmos pueden perpetuar estereotipos y desigualdades, los estudiantes comprendieron la importancia de cuestionar la objetividad de los sistemas automatizados y de ser conscientes de los datos y estructuras que los sustentan.

Otro aspecto fundamental fue el desarrollo de una conciencia ética. La exposición a dilemas éticos reales relacionados con la IA permitió a los estudiantes entender la complejidad de las decisiones que involucran el uso de estas tecnologías. Reflexionaron sobre la importancia de

considerar principios fundamentales como los derechos humanos, la privacidad y la justicia social en el desarrollo, implementación y regulación de sistemas de IA. Esta toma de conciencia ética resultó clave para sensibilizarlos sobre la necesidad de equilibrar el avance tecnológico con el respeto a los valores humanos y sociales.

La actividad también contribuyó significativamente al desarrollo de competencias digitales. La integración de herramientas digitales y el análisis de tecnologías avanzadas fortalecieron las habilidades de alfabetización digital de los estudiantes, competencias fundamentales para desenvolverse con éxito en entornos profesionales y académicos. Al interactuar con recursos digitales y analizar casos prácticos, los participantes adquirieron conocimientos aplicables a diferentes contextos, preparándolos para afrontar los desafíos de un entorno cada vez más digitalizado.

Además, se observó un notable fortalecimiento de la participación activa. Los debates guiados y las dinámicas de grupo promovieron la implicación de todos los estudiantes, quienes mejoraron sus habilidades de comunicación, argumentación y trabajo en equipo. La diversidad de opiniones y la construcción colectiva del conocimiento enriquecieron el aprendizaje, fomentando un ambiente de diálogo respetuoso donde cada voz fue escuchada y valorada. Este proceso colaborativo no solo impulsó la participación, sino que también permitió a los estudiantes desarrollar empatía y comprensión hacia diferentes perspectivas.

La actividad también fue un espacio propicio para la estimulación de la creatividad. Al enfrentarse al reto de diseñar propuestas para un uso más ético de la IA, los estudiantes pusieron en práctica su capacidad de innovación, proponiendo soluciones originales y viables. Esta combinación de pensamiento crítico y creatividad resultó fundamental para generar ideas que no solo respondieran a los dilemas éticos planteados, sino que también fueran aplicables en contextos reales. La creatividad emergió como un recurso esencial para abordar problemas complejos desde nuevas perspectivas y para imaginar futuros más justos y sostenibles.

Por último, la actividad preparó a los estudiantes para enfrentar los desafíos del mundo profesional actual. Esta preparación para la realidad profesional fue clave, ya que los estudiantes adquirieron herramientas conceptuales y prácticas para abordar retos relacionados con la ética tecnológica. Comprendieron que, como futuros profesionales, deberán enfrentarse a decisiones complejas donde la tecnología y los valores éticos entran en juego. Esta experiencia formativa los impulsó a asumir un rol activo como agentes de cambio, capaces de promover el desarrollo y la implementación responsable de tecnologías emergentes en sus respectivos campos de acción.

Por lo tanto, esta experiencia educativa no solo enriqueció el conocimiento teórico de los estudiantes sobre la inteligencia artificial y sus dilemas éticos, sino que también fortaleció habilidades prácticas esenciales para su desarrollo integral. La actividad demostró que la

formación académica debe ir más allá de la adquisición de conocimientos técnicos, incorporando la reflexión ética y el pensamiento crítico como pilares fundamentales en la educación para el siglo XXI.

DISCUSIÓN

A pesar de los resultados positivos obtenidos, la actividad "Inteligencia artificial y derechos humanos: dilemas éticos" presentó algunas limitaciones que deben ser consideradas para optimizar su impacto en futuras implementaciones. Una de las principales dificultades observadas fue la complejidad de algunos conceptos relacionados con la inteligencia artificial y la ética, lo cual generó cierta confusión en algunos estudiantes. Temas como los sesgos algorítmicos, la privacidad de los datos o la autoría de obras generadas por IA resultaron ser abstractos para aquellos que no contaban con conocimientos previos en estas áreas. Esto limitó la profundidad de los debates y afectó la participación activa de ciertos grupos.

Otra limitación detectada fue el tiempo asignado para la actividad. La amplitud de los temas tratados y la diversidad de los casos analizados requerían más tiempo para permitir una discusión más exhaustiva. El formato utilizado, aunque efectivo, resultó insuficiente para abordar en detalle cada uno de los dilemas éticos planteados. Además, algunos estudiantes manifestaron dificultades para vincular los casos prácticos con las posibles soluciones éticas, lo que evidenció la necesidad de reforzar la conexión entre teoría y práctica.

Asimismo, se identificó la falta de recursos complementarios adaptados a los diferentes niveles de comprensión de los estudiantes. La actividad no contempló materiales adicionales que ayudaran a contextualizar y profundizar en los temas abordados, lo cual podría haber facilitado el aprendizaje, especialmente para aquellos con menos experiencia previa en temas tecnológicos y éticos.

Para abordar estas limitaciones, se proponen varias mejoras. En primer lugar, sería conveniente introducir materiales previos al desarrollo de la actividad, como lecturas introductorias, videos explicativos o infografías, que permitan a los estudiantes familiarizarse con los conceptos clave. Esto ayudaría a nivelar los conocimientos y garantizar que todos los participantes partan de una base común para el análisis y la reflexión.

En segundo lugar, se recomienda extender la duración de la actividad o dividirla en varias sesiones. Esto permitiría un análisis más profundo de los casos presentados y ofrecería más tiempo para el debate, la reflexión y la elaboración de propuestas. Alternar dinámicas de trabajo individuales y grupales también podría fomentar una participación más activa y equitativa.

Además, incorporar la participación de expertos invitados en ética tecnológica o desarrollo de IA podría enriquecer los debates, aportando perspectivas profesionales y actualizadas. La

intervención de estos especialistas permitiría resolver dudas, clarificar conceptos complejos y proporcionar ejemplos concretos de cómo se abordan estos dilemas en el mundo real.

Otra propuesta de mejora consiste en diseñar actividades prácticas complementarias, donde los estudiantes puedan aplicar sus conocimientos en la creación de proyectos o soluciones innovadoras. Esto contribuiría a consolidar el aprendizaje y a vincular la teoría con la práctica.

Finalmente, sería beneficioso implementar sistemas de evaluación más detallados que permitan valorar no solo la participación en debates, sino también la capacidad de los estudiantes para aplicar principios éticos a situaciones concretas. Esto podría incluir la entrega de informes reflexivos o presentaciones orales donde expongan sus análisis y propuestas.

En conjunto, estas mejoras buscan optimizar la actividad para garantizar un aprendizaje más profundo, inclusivo y significativo, preparando a los estudiantes para enfrentar de manera crítica y responsable los desafíos éticos asociados con el uso de la inteligencia artificial.

CONCLUSIONES

El conocimiento de la legislación sobre IA y Derechos Humanos debe incorporarse en los distintos grados de la enseñanza universitaria, dada la importancia y la rapidez con la que la IA se está haciendo presente en nuestra vida diaria. La integración de la inteligencia artificial (IA) en el ámbito educativo plantea un escenario complejo, repleto de oportunidades, pero también de profundos dilemas éticos y profesionales. A lo largo de este capítulo se ha evidenciado que el avance tecnológico ha transformado los procesos de enseñanza y aprendizaje, facilitando ciertas tareas y mejorando la eficiencia. Sin embargo, también ha despertado preocupaciones legítimas sobre la calidad de la educación, la pérdida de matices culturales y pedagógicos, así como la precarización laboral de algunos profesionales de la docencia o del desarrollo de contenidos.

El análisis de la diferencia entre un problema y un dilema ha sido clave para entender la naturaleza de este desafío. Mientras que un problema puede resolverse mediante soluciones técnicas o metodológicas, un dilema requiere tomar decisiones que involucren principios éticos, valores humanos y consecuencias sociales. La aplicación de la IA en la educación no es simplemente un problema técnico que puede solucionarse ajustando algoritmos; es un dilema ético que exige evaluar hasta qué punto es moralmente aceptable sustituir el trabajo creativo y culturalmente sensible de docentes y pedagogos por sistemas automatizados.

Este dilema ético implica sopesar los beneficios indiscutibles de la IA, como la rapidez y la reducción de costes, frente a los riesgos de deshumanizar el proceso educativo y comprometer la calidad cultural, lingüística y afectiva de la formación. La postura de muchas organizaciones y asociaciones educativas refleja esta preocupación, al advertir sobre los riesgos de una

adopción indiscriminada de la IA sin regulaciones claras ni salvaguardas éticas. Es fundamental establecer límites en el uso de estas tecnologías y garantizar que las decisiones relacionadas con su implementación sean inclusivas, transparentes y centradas en el bienestar de los profesionales y la calidad de la enseñanza.

La experiencia práctica llevada a cabo en el aula, mediante la actividad “Inteligencia artificial y derechos humanos: dilemas éticos”, ha permitido trasladar este debate al contexto formativo, ofreciendo a los estudiantes la oportunidad de explorar, analizar y reflexionar sobre estos desafíos. La actividad no solo les proporcionó conocimientos sobre el funcionamiento de la IA y sus aplicaciones, sino que también fomentó habilidades fundamentales como el pensamiento crítico, la capacidad de análisis ético, la creatividad y el trabajo colaborativo. Los estudiantes comprendieron de manera tangible cómo la IA puede influir en distintos ámbitos de la sociedad y la necesidad de desarrollar soluciones éticas para enfrentar sus retos.

Ante este panorama, es imprescindible avanzar hacia un equilibrio responsable entre el uso de la IA y la protección del trabajo humano. Esto implica no solo adoptar tecnologías de manera crítica y reflexiva, sino también promover marcos regulatorios sólidos, transparentes y justos que protejan los derechos laborales y garanticen la calidad de la educación. Las instituciones académicas, las empresas tecnológicas y los órganos reguladores deben colaborar activamente para encontrar soluciones que respeten tanto el avance tecnológico como los valores humanos fundamentales.

La IA, utilizada de forma responsable y ética, puede convertirse en una herramienta valiosa que complemente el trabajo humano, liberando a los docentes de tareas repetitivas y permitiéndoles centrarse en aspectos más creativos y complejos de su labor. Sin embargo, esto solo será posible si se reconoce la importancia de la supervisión humana y se establecen límites claros para evitar la deshumanización de los procesos pedagógicos.

En conclusión, el desafío no radica en detener el avance de la IA, sino en aprender a convivir con ella de forma ética, responsable y sostenible. Solo a través de un diálogo abierto y colaborativo entre todos los actores implicados será posible construir un futuro donde la tecnología y el trabajo humano coexistan en equilibrio, garantizando la calidad educativa y protegiendo los derechos de los profesionales. La educación desempeña un papel esencial en este proceso, ya que formar a ciudadanos críticos, informados y éticamente responsables será clave para enfrentar los dilemas que plantea el desarrollo tecnológico.

Bibliografía

- Al-Zahrani, A. M., & Alasmari, T. M. (2024). Exploring the impact of artificial intelligence on higher education: The dynamics of ethical, social, and educational implications. *Humanities and Social Sciences Communications*, 11(1), 1-12.
- Aparicio-Gómez, O.-Y., & Gallego, M. A. C. (2024). Desafíos éticos de la Inteligencia Artificial en la personalización del aprendizaje. *Revista Interamericana de Investigación Educación y Pedagogía RIIEP*, 17(2), 377-392.
- Bohorques Marchori, G. (2014). *Ética y desarrollo profesional*.
- Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. 77-91.
- Cukurova, M. (2024). The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. *British Journal of Educational Technology*.
- Erik, B., & Andrew, M. (2017). The business of artificial intelligence: What it can—And cannot—Do for your organization. *Harvard Business Review Digital Articles*, 7, 3-11.
- Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.
- Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Mit Press.
- Gómez, W. O. A. (2023). La inteligencia artificial y su incidencia en la educación: Transformando el aprendizaje para el siglo XXI. *Revista internacional de pedagogía e innovación educativa*, 3(2), 217-230.
- Holmes, W. (2019). Artificial intelligence in education: Promises and implications for teaching and learning. *Center for Curriculum Redesign*.
- IEEE. (2019). *ETHICALLY ALIGNED DESIGN*.
- Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and vascular neurology*, 2(4).
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9), 389-399.
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed. *An argument for AI in Education*, 18.

Miao, F., Holmes, W., Huang, R., & Zhang, H. (2023). *AI and education: Guidance for policy-makers-UNESCO Digital Library*.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4), 2141-2168.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. En *Algorithms of oppression*. New York university press.

Parlamento Europeo y del Consejo. (2024, junio 13). *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo*. Diario Oficial de la Unión Europea.

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, 2(1), 31-42.

Seldon, A., Abidoye, O., & Metcalf, T. (2020). *The Fourth Education Revolution Reconsidered: Will Artificial Intelligence Enrich Or Diminish Humanity?* Legend Press Ltd.

Selwyn, N. (2019). *Should robots replace teachers?: AI and the future of education*. John Wiley & Sons.

Shamim, S., Cang, S., Yu, H., & Li, Y. (2016). *Management approaches for Industry 4.0: A human resource management perspective*. 5309-5316.

Solove, D. J. (2005). A taxonomy of privacy. *U. Pa. l. Rev.*, 154, 477.

Stahl, B. C. (2021). *Artificial intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies*. Springer Nature.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44-56.

UNESCO, C. (2021). *Recommendation on the ethics of artificial intelligence*.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications*, 11(1), 1-10.

Luis Demano. La inteligencia artificial generativa y el Artículo 27 de la Declaración Universal de los Derechos Humanos

Luis Demano

En representación de la Associació de Professionals de la Il·lustració Valenciana (APIV)

Declaración Universal de los Derechos Humanos (DUDH, 1948) Artículo 27.

1. *Toda persona tiene derecho a tomar parte libremente en la vida cultural de la comunidad, a gozar de las artes y a participar en el progreso científico y en los beneficios que de él resulten.*

2. *Toda persona tiene derecho a la protección de los intereses morales y materiales que le correspondan por razón de las producciones científicas, literarias o artísticas de que sea autora.*

Propuesta basada en la investigación en curso sobre el impacto de la inteligencia artificial generativa en el sector cultural y la vulneración de los derechos de propiedad intelectual de todas las personas profesionales por parte de las empresas de inteligencia artificial generativa, desarrollada gracias a una beca del Ministerio de Cultura del Gobierno de España.

El Artículo 27 de la Declaración Universal de los Derechos Humanos y su vulneración por parte de las empresas de inteligencia artificial generativa se sustenta en el hecho, demostrable y admitido públicamente por las propias empresas tecnológicas, del robo masivo —mediante la técnica del *web scraping*— y posterior utilización ilícita de obras protegidas por derechos de autor para el entrenamiento de sus modelos de inteligencia artificial generativa y su posterior comercialización.

El primer estudio mundial que mide el impacto económico de la IA generativa en las industrias audiovisual y musical, encargado por la CISAC —Confederación Internacional de Sociedades de Autores y Compositores, que representa a más de 5 millones de creadores—, confirma que «la IA generativa enriquecerá a las empresas tecnológicas y comprometerá sustancialmente los ingresos de los creadores humanos en los próximos años».

Más allá de los beneficios potenciales de la innovación tecnológica —que sin duda existen—, debemos prevenir sus efectos negativos, más acentuados en sectores profesionales atravesados por una precariedad laboral endémica, como es el caso del sector cultural.

En el verano de 2022 la empresa tecnológica OpenAI lanzó al mercado su nuevo modelo de Inteligencia Artificial generativa DALL-E 2, cuyos resultados en la generación de textos e imágenes, a partir de datos previos, comenzaron a provocar muchas preguntas entre las personas profesionales del sector cultural. Preguntas que, gracias al esfuerzo de la comunidad artística organizada internacionalmente, pronto arrojarían muchas preocupantes respuestas.

En septiembre de ese mismo año, Lapine —artista digital residente en San Francisco—, haría un interesante descubrimiento mientras utilizaba la herramienta *Have I Been Trained*,¹ que permite a los artistas comprobar si su trabajo es utilizado para entrenar modelos de IA generativa. *Have I Been Trained* fue creada por los artistas Mat Dryhurst y Holly Herndon, cuyo trabajo los llevó a explorar cómo las empresas tecnológicas de IA se apropian de las obras protegidas por derechos de autor. Cuando Lapine utilizó esta herramienta para escanear la principal y más importante base de datos utilizada por las empresas tecnológicas, LAION 5B, lo primero que encontró fue una imagen de su propia cara. A través del rastreo de esta imagen, pudo descubrir que se trataba de una de las fotografías tomadas por su médico durante un tratamiento y cuyo uso —bajo firma— se limitaba únicamente a su historial clínico, protegido por las leyes de protección de datos. Su médico había fallecido en 2018, así que no podría responder sobre cómo su imagen médica privada había acabado digitalizada y online, para más tarde pasar a formar parte de la principal base de datos utilizada por las empresas tecnológicas de IA generativa para el entrenamiento de sus modelos. Si algo de naturaleza tan íntima había acabado sin su consentimiento en manos de una empresa privada con ánimo de lucro, qué estaría sucediendo con su obra artística.

Pocos meses más tarde, fue el artista Steven Zapata en su viralizado video *The End of Art: An Argument Against Image AIs*² quien continuó haciéndose preguntas y marcando un punto de inflexión dentro de la comunidad artística y la industria cultural, provocando así una generalizada toma de conciencia del verdadero problema al que se enfrentaban las personas profesionales del sector y las dimensiones abismales del mismo. En su video, Zapata analizaba, además del impacto ético de la IA generativa, algunas de las claves técnicas de esta tecnología. Entre las más importantes, se encontraba la creación de las bases de datos a partir de las cuales los modelos de IA generativa se habían entrenado. Era muy sencillo de entender: para que un modelo pueda generar una imagen necesita disponer de una base de datos previa. En el caso de Stable Diffusion y Google Image, su base de datos era LAION 5B, una colección de 5.000 millones de imágenes y textos descriptivos que en una gran proporción estaban protegido por derechos de autoría.

LAION 5B es creada por LAION, una organización alemana sin ánimo de lucro, mientras que las imágenes y los textos son rastreados mediante la práctica del *web scraping* por otra organización sin ánimo de lucro llamada *CommonCrawl*. Dependiendo de dónde se encuentren estas empresas, al estar registradas como organizaciones sin ánimo de lucro y con fines de investigación, gozan de exenciones legales. No serán declaradas culpables por recopilar y utilizar datos que, de otro modo, sí que estarían protegidos por derechos de autor. ¿Y quién financió LAION para crear su base de datos? Una empresa privada con ánimo de lucro: Stability AI, dueña del popular modelo de IA generativa Stable Diffusion. De este modo, a través de una empresa sin ánimo de lucro, con fines de investigación y creada en un país donde disfruta de exenciones legales, se canaliza posteriormente todas las obras protegidas por derechos de autor a empresas con ánimo de lucro como Stability AI, una empresa que llegó a alcanzar un valor de mercado de mil millones de dólares gracias a su producto estrella Stable Diffusion. O a OpenAI, cuyo principal socio comercial y proveedor de infraestructura en la nube, Microsoft, es ya la segunda empresa de la historia —Nvidia fue la primera— en alcanzar una valoración bursátil de 4 billones de dólares.

«En el caso de DALL-E», continuaba Zapata en su video, «su modelo de evasión de responsabilidades es el llamado OpenAI, formado por la corporación con ánimo de lucro OpenAI LP y su empresa matriz, la organización sin ánimo de lucro OpenAI Inc». Quienes meses más tarde lanzarían otro famoso competidor al mercado: ChatGPT (*Chat Generative Pre-Trained Transformer*). En el caso de Midjourney, otra popular inteligencia artificial generativa, ni tan siquiera han querido hacer pública la información sobre el origen de su base de datos. Pero de nuevo, gracias al esfuerzo de la comunidad artística organizada internacionalmente, se acabó descubriendo cómo funciona su modelo de negocio.

THE LOST PENNY FILES

En agosto de 2023, una persona llamada Brian Penny publicó en la red social Twitter algunas conversaciones del canal de Discord de Midjourney —del que era miembro— posteriormente apodadas bajo el nombre de The Lost Penny Files. Su contenido esclarecía los primeros pasos del desarrollo de Midjourney y Stable Diffusion, ya que tanto David Holz (fundador de la primera) como Emad Mostaque (CEO de Stability AI en aquel momento) aparecían activamente en las conversaciones. Dos temas principales salieron a la luz. El primero, constata la existencia de un «canal oculto (*war room*) donde el personal de Midjourney y sus primeros usuarios generaron imágenes perturbadoras y potencialmente ilegales relacionadas con violencia, drogas, política y abuso sexual infantil (CSAM) no apto para el trabajo (NSFW)», publica el medio Luddite Pro.³ «¿Podrías enviármelas?» pregunta David Holz. Así es como descubre, con sus propios ojos, lo que es capaz de generar la tecnología que están ayudando a crear. «Si un modelo sabe qué es un niño y qué es la pornografía...» comenta un empleado de Midjourney en clara referencia a la capacidad generativa —mediante la combinación de referentes

previos— de los modelos de IA que están entrenando. «Canal NSFW, genial, publicaré allí si quiero probar algo atrevido en la *war room*. Voy a por todas» comenta otro usuario (Somnai) con pasmosa frivolidad.

El segundo tema principal del canal de Discord de Midjourney es la propiedad intelectual, que a su vez podría dividirse en tres momentos claves de la conversación.

En el primero, Emad Mostaque presenta por primera vez —antes de su lanzamiento en marzo de 2022— LAION 5B, el mayor conjunto de datos hasta el momento, que incluía más de 5 mil millones de imágenes sustraídas de internet, entre las que se encontraban millones de obras de artistas protegidas por derechos de autor. La reacción de David Holz no se hizo esperar, anunciando que Midjourney también está en proceso de recolección de datos para el entrenamiento de sus modelos, motivando la siguiente pregunta de uno de los miembros del canal: «¿Planeas entrenar en Artstation o te alejarás de posibles problemas legales?». «Parece que tendremos acuerdos comerciales para conjuntos de datos especiales» contesta Holz. Casi de inmediato, es Moustaque quien lanza su oferta, celebrada con entusiasmo por el resto: «Deberíamos reunir a todos, ya que estamos alineando múltiples modelos. Planificarlo todo tal vez sea la mejor opción».

En lo respecta a la mencionada plataforma Artstation —una de las comunidades de artistas online más importante—, el posicionamiento ambiguo por parte de Tim Sweeney (CEO de Epic Games, empresa matriz de Artstation) destapó recelos tras aparecer mencionada por David Holz en su conversación en Discord. Aunque Sweeney se mostró, en principio, contrario al uso de las obras alojadas en su plataforma para el entrenamiento de la IA generativa, su opinión sobre el respeto a la propiedad intelectual no tardó en cambiar: «El aprendizaje humano mediante la lectura de libros con derechos de autor y la observación de obras de arte no es robar. ¿Por qué asumiríamos que el aprendizaje automático lo es?».

El segundo momento clave de la conversación sobre propiedad intelectual en el canal de Discord de Midjourney fue la aparición de Brad Templeton, expresidente de la Electronic Frontier Foundation (EFF), en abril de 2022. Enseguida su conversación con David Holz se enfocó hacia el potencial comercial de Midjourney, «¿Es esto un juego o una herramienta productiva?», pregunta Templeton. «Si millones de personas quieren jugar con esto, su fuerza cultural arrastrará todo lo demás», vaticina Holz. «Deberíamos evitar el uso comercial en la primera etapa», añade. Templeton lo tiene claro: «Vas a tener un problema de derechos de autor». Varios miembros del canal responden de inmediato e improvisan justificaciones con la intención de disuadir a Templeton. El primero en reaccionar es Holz: «No se utiliza ninguna base de datos de imágenes en el proceso de creación de una imagen», algo que se ha demostrado rotundamente falso. Otro miembro del canal, trabajador de Midjourney, no duda en reforzar a su jefe: «Estos modelos funcionan sin internet, es como una caja de Pandora en

tu máquina». Para entonces, LAION 5B ya había arrasado por primera vez —a través de Common Crawl, como ya hemos visto— con todo el contenido de internet para entrenar a los modelos. Aún así, Holz redundaba en la mentira: «Sí, exactamente. Sin internet. Sin base de datos. Simplemente [el modelo] “sabe” todas estas cosas».

Brad Templeton, cada vez más desconcertado, trata de entender a sus interlocutores: «Entonces, cuando le digo que dibuje a una persona concreta, ¿cómo sabe dibujar a esa persona? Cuando le pregunto por el estilo de un artista, ¿cómo puede reproducir ese estilo?». «Sabe cómo son y ha visto estilos» insiste Holz; «Entrenando con cientos de millones de imágenes de internet» añade su empleado; «Lo aprendió y está codificado en miles de millones de pesos neuronales indescifrables, al igual que nuestros cerebros», «Básicamente como se entrena a un artista humano». Desde este momento, la falacia de la metáfora computacional —atribuir cualidades humanas a una máquina— empezó a utilizarse como excusa recurrente para eludir responsabilidades. Un argumento que las redes sociales extenderían como una mancha de aceite entre los usuarios de esta tecnología.

El tercer momento clave de la conversación fue una nueva aparición de Somnai —aquel usuario motivado con la *war room*—, figura fundamental para comprender el proceso de creación de la tecnología que arrasaría con los derechos de millones de artistas y profesionales de la industria cultural.

Dos meses antes de que Templeton entrara en escena y formulase la pregunta citada anteriormente sobre «el estilo de un artista», el tema en cuestión ya sobrevolaba las conversaciones del canal, ganando progresivamente protagonismo hasta convertirse en el punto clave del debate sobre propiedad intelectual. «Lo siento, ¿cómo se usan los estilos en los prompts?» preguntaba un usuario despistado, «Se trata principalmente de nombres de artistas: 4.000 artistas, 1.000 estilos» responde David Holz en el hilo de la conversación. Así que sólo tenías que incluir el nombre del artista en el texto del prompt y te generaría una imagen simulando su estilo. «Ojalá hubiera una lista grande de artistas contemporáneos» comentaba otro usuario. Y aquí es donde aparece de nuevo el usuario Somnai: «Tengo unos 16.000 artistas, ¿demasiados?». Holz no lo dudó ni un instante: «Para nada, simplemente colócalos en la sección de propuestas y pon “nombre del artista” al lado».

Varias investigaciones, realizadas por los miembros más implicados de la comunidad artística, descubrieron que la persona detrás del usuario Somnai era, en realidad, el diseñador y artista 3D llamado Maxwell Ingham. También responsable del desarrollo de la IA generativa de imágenes Disco Diffusion. Una vez desvelado su anonimato, lo que más sorprendió a quienes lo investigaron fue un ligero detalle que, en un principio, les pasó desapercibido: Ingham no incluyó en ningún momento su propio nombre en la lista de los 16.000. El único Maxwell incluido era el famoso animador y guionista estadounidense Maxwell Atoms. ¿No se

consideraba tal vez lo suficientemente talentoso como para compartir espacio con los miles de artistas que sí incluyó en su lista? ¿o respondía su omisión a no querer que Midjourney entrenase con su trabajo? Nunca lo sabremos. Lo que sí podemos confirmar es que, un mes después de ofrecer en el canal el contenido de la lista de los 16.000, fue contratado en Midjourney como Senior Software Developer.

Muchos otros intentos por parte de la comunidad artística de conseguir una información tan detallada como la revelada en el canal de Discord de Midjourney, se han visto frustrados por la opacidad con la que han operado los dueños de los modelos fundacionales de IA generativa. Un flujo subterráneo de información, evidencias hurtadas a la gran mayoría con las que ya no podremos contar para profundizar en la «otra» historia de esta tecnología. Aún así, lo que sí ha quedado demostrado sin ningún tipo de dudas es que sin los artistas y su trabajo sería imposible la existencia y comercialización de los actuales modelos de IA generativa, pero no a la inversa. Y las empresas y sus dueños lo saben. Nick Clegg —ex viceprimer ministro británico y ex vicepresidente de Asuntos Globales y Comunicación de Meta—, es otra de las figuras políticas vinculadas al sector tecnológico que fue tajante respecto a esta dependencia. Coincidiendo con las sesiones en el Parlamento de Reino Unido para la regulación de la IA generativa, le preguntaron sobre el derecho de la comunidad artística a decidir si sus obras pueden utilizarse para entrenar los modelos de IA: «Desconozco cómo se podría preguntar previamente a todo el mundo. No veo cómo podría funcionar», comentaba contrariado, «Y por cierto, si lo hicieran en Gran Bretaña y nadie más lo hiciera, prácticamente acabarían con la industria de la IA en este país de la noche a la mañana».⁴

DEMANDAS COLECTIVAS

Otra base de datos revelada gracias a una investigación del diario estadounidense The Washington Post, ha sido la C4 (*Colossal Clean Crawled Corpus*) de Google. Recopilada también a través de la organización sin ánimo de lucro CommonCrawl —como LAION 5B— está formada a partir del contenido de «sitios web de industrias como el periodismo, el entretenimiento, el desarrollo de software, la medicina y la creación de contenidos, lo que ayuda a explicar por qué estos campos pueden verse amenazados por la nueva ola de inteligencia artificial. Los tres sitios más importantes fueron patents.google.com, que contiene texto de patentes emitidas en todo el mundo; wikipedia.org, la enciclopedia gratuita en línea; y scribd.com, una biblioteca digital de acceso restringido por suscripción. También ocupa un lugar destacado en la lista b-ok.org, un famoso mercado de libros electrónicos pirateados confiscado posteriormente por el Departamento de Justicia de Estados Unidos. Al menos otros 27 sitios identificados por el gobierno de EEUU como mercados de piratería y falsificaciones están presentes en el conjunto de datos C4».⁵ Curiosamente, el símbolo de *copyright* señalando que una obra está registrada como propiedad intelectual, aparece más de 200 millones de veces en el conjunto de datos C4.

Una vez revelada la información sobre cómo se construyeron las bases de datos necesarias para el entrenamiento de las inteligencias artificiales generativas, las empresas tecnológicas negaron en rotundo la evidencia. No mucho más tarde, presionadas por las incesantes campañas de denuncia de la comunidad artística, terminaron admitiendo su responsabilidad por la vulneración de los derechos de autor.

La primera en hacerlo fue Stability AI, cuando dos meses después de la salida de Stable Diffusion lanzó al mercado Dance Diffusion, su modelo de audio. En su web podía leerse el siguiente texto:

«Dance Diffusion también se basa en conjuntos de datos compuestos en su totalidad por muestras de música y audio libres de derechos de autor y proporcionadas voluntariamente. Dado que los modelos de difusión son propensos a la memorización y al sobreajuste, la publicación de un modelo entrenado con datos protegidos por derechos de autor podría dar lugar a problemas legales. Para honrar la propiedad intelectual de los artistas y, al mismo tiempo, cumplir en la medida de lo posible con las normas de derechos de autor, a menudo estrictas, de la industria musical, era imprescindible mantener cualquier tipo de material protegido por derechos de autor fuera de los datos de entrenamiento».

En nuestro segundo caso a destacar, la declaración realizada por la empresa OpenAI al medio The Guardian en enero de 2024 es meridiana: «Es imposible crear herramientas de IA como ChatGPT sin material con derechos de autor».⁶

Ante estas y otras evidencias posteriores, la respuesta legal de la comunidad artística y las empresas afectadas por el rastreo masivo de datos y la vulneración de propiedad intelectual no se hizo esperar. Las primeras en organizarse y reaccionar fueron las ilustradoras Kelly McKernan, Karla Ortiz, and Sarah Andersen, con la ayuda de la Concept Art Association de EEUU. Su demanda colectiva implicaba a Stability AI (Stable Diffusion), Midjourney y la web DevianArt por entrenar a su IAG, Dreamup, con las obras propiedad de las artistas alojadas en su web. La denuncia incide en la posibilidad de incluir el propio nombre de las artistas en la descripción de texto (*prompt*) para poder replicar el estilo de las artistas y generar contenido que perjudique su imagen profesional. Un año más tarde, Getty Images, la empresa más importante de imágenes de stock, presentó su propia demanda contra Stability AI en Londres.

Algunos meses después, siete nuevos artistas se sumaron a la primera demanda: Hawke Southworth, Grzegorz Rutkowski, Gregory Manchess, Gerald Brom, Jingna Zhang, Julia Kaye y Adam Ellis. Incluyendo además a una nueva empresa en la denuncia: Runway AI, la empresa que entrenó al modelo Stable Diffusion 1.5. Uno de los elementos más destacable en la nueva ampliación de la denuncia fue el artículo académico del investigador de Google DeepMind, Nicholas Carlini, donde él y su equipo demuestra que las IAGs se entrenan intencionadamente para reconstruir las obras que forman su base de datos:

«Los modelos de difusión de imágenes como DALL-E 2, Imagen y Stable Diffusion han atraído una gran atención debido a su capacidad para generar imágenes sintéticas de alta calidad. En este trabajo, demostramos que los modelos de difusión memorizan imágenes individuales a partir de sus datos de entrenamiento y las emiten en el momento de la generación. Mediante un proceso de generación y filtrado, extraemos más de mil ejemplos de entrenamiento de los modelos más avanzados, desde fotografías de personas hasta logotipos de marcas comerciales».

Es más, en las imágenes generadas por estos primeros modelos podía observarse, en una esquina, mimetizada sin disimulo, la firma del artista original cuya obra había sido utilizada como base para la reproducción. En las versiones posteriores de los modelos, esta impronta humana, convertida en prueba fehaciente del uso ilícito, sería eliminada del resultado final para crear en sus usuarios la falsa ilusión de estar participando de un proceso creativo real.

En otro caso de interés, el experto en inteligencia artificial de la Universidad de New York, Gary Marcus, en colaboración con el artista conceptual Reid Southen, logró desestabilizar la estrategia de omisión de responsabilidades que la industria tecnológica había construido con tanto empeño. En su investigación publicada bajo el nombre La IA generativa tiene un problema de plagio visual,⁷ se hacían la siguiente pregunta: ¿Es posible inducir a los modelos de generación de imágenes a producir resultados plagiados a partir de materiales protegidos por derechos de autor? Para sorpresa de nadie, la respuesta era claramente afirmativa. El estudio demostraba que los resultados generados por los modelos eran una copia exacta de las imágenes originales que se utilizaban para entrenarlos. Fotogramas de películas de Marvel, Warner Bros y Disney, escenas de famosos videojuegos como The Last of Us, el universo Star Wars, o la familia Simpson eran duplicados de forma algorítmica. Incluso sin necesidad de hacer referencia directa al objeto del plagio, los modelos arrojaban calcos de la obra original a capricho del consumidor. Con tan solo incluir en el texto del *prompt* la expresión *screencrap* (captura de pantalla) «de una película popular», podías obtener una versión de cualquier personaje de la cultura popular.

La amplia difusión y viralidad conseguida por la investigación de Marcus y Southen la convirtió en un valioso referente de la lucha por la defensa de los derechos de autor, tanto para la comunidad artística como para la industria cultural más corporativa. Tanto es así, que Disney y NBCUniversal la incluyeron como prueba —como con la ya citada investigación de Carlini— para fundamentar la primera gran demanda federal de estas dos empresas de Hollywood contra Midjourney por violación de derechos de autor en 2025. Según declaraciones de ambas compañías, Midjourney es un «fondo infinito de plagio».

Todo lo anterior, sumado al informe publicado por el Stanford Internet Observatory,⁸ fue lo que realmente hizo saltar las alarmas sobre el contenido de las bases de datos de

entrenamiento. Según la investigación que reflejaba el informe, difundido por varios medios, LAION 5B contenía ingente material ilegal de abuso sexual infantil (CSAM), con el que Stable Diffusion habría sido entrenado. Advirtiendo, además, que la cantidad de material pedófilo podría ser mucho mayor, «dado que nuestra evaluación se centró en una pequeña porción del conjunto de miles de millones de datos». Otros modelos como Midjourney, cuya transparencia en el entrenamiento de los modelos es nula, podrían contener también este tipo de material.

Como es evidente, la lucha de los y las artistas en favor de la regulación y la estricta supervisión ética de los modelos de Inteligencia Artificial generativa, va más allá de su interés, que algunos tachan de egoísmo, por defender su parcela de mercado profesional. Consta que, sin control democrático y transparencia, las empresas privadas de IA seguirán utilizando cualquier recurso que les dé ventaja sobre el resto de sus competidores.

REGLAMENTO EUROPEO DE INTELIGENCIA ARTIFICIAL

El 13 de junio de 2024 se publicó en el Diario Oficial de la Unión Europea el conocido como Reglamento Europeo de Inteligencia Artificial —también conocida como ley Europea de Inteligencia Artificial—, una norma pionera y de referencia internacional donde se destacan los posibles riesgos para los derechos fundamentales derivados del uso de sistemas de inteligencia artificial. Entre otros, aparecen nombrados el derecho a la dignidad humana, la libertad de expresión y de información, los derechos de los trabajadores y los derechos de propiedad intelectual.

Como se comentaba anteriormente, todos los sistemas de IA generativa actuales se crean con cantidades masivas de datos recopilados, reproducidos y almacenados para su posterior proceso de entrenamiento. Multitud de esos datos son textos, imágenes, música y otras obras intelectuales cuyos derechos de autor y de propiedad intelectual pertenecen a los autores, artistas y los legítimos titulares a los que se les hayan cedido esos derechos. Sin embargo, todos los modelos de IA generativa comercializados en la actualidad han sido entrenados sin su consentimiento previo ni, por supuesto, haberles remunerado por la cesión de su uso para un propósito comercial.

Por lo tanto, la gran mayoría de las empresas propietarias de estos modelos de IA generativa han vulnerado la normativa existente, tanto en el ámbito internacional como nacional, relativa a los derechos de autor y de propiedad intelectual que se fundamenta en dos principios básicos y universales: primero, sin autorización, no hay uso legal, salvo que exista una excepción de contexto aplicable; segundo, todo uso ha de ser remunerado.

Ambos principios están reconocidos internacionalmente y se fundamentan en el respeto a la creatividad humana de las personas y al derecho a poder vivir de su trabajo con dignidad.

Los modelos de IA generativa ponen en riesgo estos principios cuando se lanzan al mercado para su explotación comercial, generando productos —que no obras— que, en muchas ocasiones, como hemos visto anteriormente, son idénticos a aquellas obras que les sirvieron como material de entrenamiento. Se realizan, por tanto, reproducciones o transformaciones de las obras y se ponen a disposición de terceros, todo ello sin autorización ni cesión de derechos de sus titulares y sin remunerarles por su uso. Además, estos productos generados son utilizados por terceros que los ofertan en los mismos contextos, lugares, mercados o plataformas donde autores, artistas y otros titulares de derechos comercializan obras originales, compitiendo con ellos de forma desleal, desincentivando la creación artística y poniendo en riesgo su profesión y sus fuentes de ingresos.

En el marco jurídico de la Unión Europea, estas empresas propietarias de los modelos justifican sus acciones aprovechando las excepciones introducidas por la Directiva (UE) 2019/790 relativa a la minería de textos y datos, permitiendo las reproducciones y las extracciones de las obras protegidas. Sin embargo, estas excepciones no se crearon para que sirvieran para la extracción y uso ilegal, masivo e indiscriminado de obras para el entrenamiento de los modelos de IA generativa.

De hecho, esta Directiva fue aprobada en 2019, mucho antes de la aparición y comercialización de los actuales modelos de IA generativa, por lo que los legisladores que la desarrollaron no contemplaban ese riesgo. No se puede poner la venda en ausencia de herida. La función principal de la Directiva era posibilitar «el tratamiento de grandes cantidades de información para adquirir nuevos conocimientos y descubrir tendencias», fomentar la investigación y la innovación en el ámbito académico y en el sector privado, pero nunca promover el uso indiscriminado e ilegal de obras para generar, de forma masiva, otros productos que compitieran deslealmente con las obras de las que proceden.

Además, la Directiva (UE) 2019/790 introdujo otra práctica vinculada al uso de la excepción de minería de textos y datos en el sector privado, contradiciendo así los principios básicos de propiedad intelectual. Esta práctica, exige a las personas titulares de los derechos de las obras que, si no quieren que estas sean utilizadas, lo manifiesten expresamente mediante el llamado mecanismo opt-out. O sea, una vez que las obras protegidas por derechos de autor han sido utilizadas para entrenar a los modelos, los propietarios de derechos tienen la posibilidad de que sus obras sean eliminadas de las bases de datos de entrenamiento. Algo imposible de poner en práctica, dada la opacidad con la que operan las empresas tecnológicas durante todas sus fases del proceso. La introducción de este mecanismo opt-out sirve a las empresas para eludir responsabilidades legales por el uso indiscriminado y sin permiso de millones de obras para entrenar a sus modelos.

Además, la ligereza con la que el Reglamento Europeo de Inteligencia Artificial ha introducido la Directiva (UE) 2019/790 como base de regulación de la tecnología, ha derivado en graves consecuencias para los intereses de las personas titulares de derechos de autor. Según plantea la Directiva, los organismos de investigación pueden realizar minería de textos y datos de obras protegidas sin necesidad de autorización de las personas titulares de derechos, siempre que hayan accedido lícitamente a las mismas y la finalidad sea la investigación científica. Un escenario que ha provocado que en el ámbito universitario se estén desarrollando modelos de IA generativa entrenados con obras protegidas y que, amparándose en la excepción de la Directiva, los estén poniendo a disposición de empresas privadas con ánimo de lucro. La política de *Open Science* de la Unión Europea que financia estos proyectos, obliga a poner a disposición de los ciudadanos los datos, la metodología, los procesos y los resultados de su investigación.

La consecuencia directa es que las empresas del sector privado, aprovechando los modelos de IA generativa desarrollados bajo este marco, entienden que para este uso de explotación comercial no deben solicitar ninguna autorización ni cesión de derechos. Tampoco remunerar a los autores que, involuntariamente, se ven involucrados en esos mismos procesos de explotación comercial sin recibir una compensación proporcional al valor que su trabajo aporta a la tecnología implementada por las empresas. Esta interpretación de la Directiva subvierte claramente sus fundamentos y atenta directamente contra la propiedad intelectual.

Esta situación de total vulnerabilidad ante estructuras económicas con una mayor capacidad de ejercer su poder ante sectores profesionales más débiles, no sólo afecta a las personas autoras, artistas y titulares de los derechos sobre estas obras en lo relativo a su propiedad intelectual, sino que tiene también un grave impacto negativo sobre la cantidad de encargos que reciben, los ingresos derivados de ellos y, en consecuencia, en las posibilidades que tienen de vivir de su trabajo, mientras otros se benefician económicamente del mismo. Esta situación, además de desincentivar la actividad artística, afecta a muchas otras personas trabajadoras de la industria cultural que, aun no teniendo derechos de propiedad intelectual, son esenciales durante todo el proceso de creación de las obras y cuyas labores fundamentales están siendo sustituidas por los modelos de IA generativa.

Ante esta situación de desamparo, deben protegerse el derecho a la libertad de expresión y de creación de las personas autoras. Incluyendo el derecho al trabajo digno de todas las personas trabajadoras de la cultura. Si queremos que los seres humanos sigan creando y expresándose, también a través de su trabajo, debemos establecer medidas que regulen los modelos de IA generativa que comercializan productos generados compitiendo deslealmente en el mismo contexto profesional.

El respeto a los derechos de autor y la propiedad intelectual para proteger a las personas trabajadoras del sector cultural deben incluir los siguientes puntos básicos:

· Autorización previa.

Sin permiso, autorización o cesión de derechos previos las obras artísticas y literarias no se pueden utilizar ni en la fase de entrenamiento ni en cualquiera de las otras fases de desarrollo y puesta en marcha de los modelos de inteligencia artificial. El mecanismo de reserva expresa de los derechos sobre las obras por parte de sus titulares infringe el principio básico de la propiedad intelectual que es la autorización previa.

· Remuneración por el uso legal de las obras.

Cualquier uso que sea autorizado en el ámbito de la inteligencia artificial debe ser remunerado conforme con lo establecido en la normativa de propiedad intelectual.

- Retirada del mercado de los modelos ilegales de IA generativa.
- Indemnizando a las personas afectadas por el uso ilegal de obras realizado desde la aparición de los modelos, habiendo causado un perjuicio económico y moral a las personas titulares de derechos.
- Las excepciones de minería de textos y datos no deben ser aplicables a la IA generativa.
- Estas excepciones establecidas en la Directiva (UE) 2019/790 no pueden servir para el uso de obras artísticas en el entrenamiento de los modelos, ya que no fueron pensadas ni creadas para ello.
- Los productos generados por la IA no pueden ser considerados obras de creación intelectual ni pueden estar protegidos por la normativa de derechos de autor y de propiedad intelectual. Los productos (imágenes, textos, música, etc.) que sean resultado, total o parcial, de los modelos de IA generativa no deben considerarse obras de creación intelectual y, por tanto, ni las empresas propietarias de dichos modelos ni los terceros que los utilicen tendrán ningún derecho de autor ni de propiedad intelectual sobre ellos.
- Los productos generados por la IA deben ser identificados como tal y no deben presentarse a concursos ni premios artísticos de carácter público, ni utilizarse para competir en licitaciones públicas con profesionales de la cultura.
- Obligación de transparencia e información accesible por parte de las empresas de IA generativa para que las personas titulares de derechos de autor puedan saber si sus obras han sido utilizadas para el entrenamiento, cualquiera de las fases de desarrollo, puesta en marcha y explotación de los modelos de IA generativa. Las empresas propietarias de estos modelos deben informar de forma específica sobre las obras concretas que han sido utilizadas, no siendo suficiente una mera descripción de un corpus o conjunto de obras sin especificación de aquellas singularmente incluidas.
- Considerar como ilegales las cláusulas de consentimiento tácito para el uso de las obras para el entrenamiento de la IA por la mera utilización de plataformas y redes sociales.

Las condiciones generales de servicios de las principales plataformas y redes sociales instauran el consentimiento tácito de sus usuarios para que sus obras sean utilizadas en el entrenamiento y desarrollo de la IA deben anularse, puesto que las cesiones de derechos sobre las obras siempre deben reflejarse en un contrato por escrito consensuado por ambas partes.

Dado que ninguno de estos puntos básicos se aplica actualmente para proteger en el presente y evitar futuros daños a las personas profesionales del sector cultural, las consecuencias negativas ya se están visibilizando.

El primer estudio¹⁰ —de ámbito mundial— que mide el impacto económico de la IA en las industrias audiovisual y musical, encargado por la CISAC (Confederación Internacional de Sociedades de Autores y Compositores, que representa a más de 5 millones de creadores), confirma que «la IA generativa enriquecerá a las empresas tecnológicas y comprometerá sustancialmente los ingresos de los creadores humanos en los próximos años». Según el estudio, las pérdidas acumuladas de aquí al 2028 en el sector musical y audiovisual serían del 24% y 21% respectivamente. Lo que equivaldría a 22.000 millones de euros en un periodo de cinco años (10.000 millones de euros en el sector musical y 12.000 millones de euros en el sector audiovisual). Estas pérdidas, continúa el estudio, «se derivan directamente de la reproducción no autorizada de las obras de los creadores y representan una transferencia del valor económico de los creadores a las empresas de IA generativa». En unos pocos años, el contenido musical generado por IA supondrá el 20% de los ingresos de las plataformas de streaming y alrededor del 60% de los ingresos de las bibliotecas musicales. En la industria audiovisual, los traductores y adaptadores de doblajes y subtítulos perderán el 56% de sus ingresos, mientras que para los guionistas y directores supondrá entre un 15% y un 20%.

El actual marco regulador provoca las pérdidas de las personas profesionales en dos frentes: la pérdida de ingresos causada por el uso no autorizado de sus obras por modelos de IA generativa sin ningún tipo de remuneración; y el reemplazo de sus fuentes tradicionales de ingresos debido al efecto de sustitución de los productos generados por IA, que ya están compitiendo con las obras creadas por humanos.

Para revertir la situación actual se debe reforzar —además del ya mencionado Reglamento Europeo de Inteligencia Artificial— la redistribución de recursos por parte del Estado, apoyando así la difusión y divulgación de las demandas del sector cultural afectado por los intereses privados de las empresas tecnológicas. Más allá de pensar en los beneficios potenciales de la innovación tecnológica, que sin duda existen, debemos prevenir sus efectos negativos, generalmente más acentuados en sectores profesionales atravesados por una inestabilidad y precariedad laboral endémica, como es el caso del sector cultural. La dignidad de las personas y la defensa de sus derechos en el desarrollo de sus actividades profesionales debe estar por

encima de cualquier supuesto avance tecnológico en el contexto de una sociedad democrática. Todavía estamos a tiempo de paliar los perjuicios ocasionados y construir una base sólida para un uso legal y ético de la inteligencia artificial generativa en el futuro.

¹ *Have I been trained* [fecha de consulta: 12 de octubre de 2025]. Disponible en: <https://haveibeen trained.com/>

² Steven Zapata. *The End of Art: An Argument Against Image AIs* [vídeo en línea]. 2022 [consulta: 12 de octubre de 2025]. Disponible en: <https://www.youtube.com/watch?v=tjSxFAGP9Ss>

³ *Luddite Pro* [fecha de consulta: 5 de septiembre de 2023]. Disponible en: <https://luddite.pro/>

⁴ SATO, Mia. *Nick Clegg says asking artists for use permission would 'kill' the AI industry*. En: *The Verge* [en línea]. 26 de mayo de 2025 [consulta: 3 de noviembre de 2025]. Disponible en: <https://www.theverge.com/news/674366/nick-clegg-uk-ai-artists-policy-letter>

⁵ SATO SCHAUL, Kevin Schaul; CHEN, Szu Yu; TIKU, Nitasha. *Inside the secret list of websites that make AI like ChatGPT sound smart*. En: *The Washington Post* [en línea]. 19 de abril de 2023 [consulta: 5 de mayo de 2023]. Disponible en: <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>

⁶ SATO MILMO, Dan. *'Impossible' to create AI tools like ChatGPT without copyrighted material, OpenAI says*. En: *The Guardian* [en línea]. 8 de enero de 2024 [consulta: 8 de enero de 2024]. Disponible en: <https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material-openai>

Bibliografía

1. Have I been trained [fecha de consulta: 12 de octubre de 2025]. Disponible en: <https://haveibeen trained.com/>
2. Steven Zapata. *The End of Art: An Argument Against Image AIs* [vídeo en línea]. 2022 [consulta: 12 de octubre de 2025]. Disponible en: <https://www.youtube.com/watch?v=tjSxFAGP9S>
3. Luddite Pro [fecha de consulta: 5 de septiembre de 2023]. Disponible en: <https://luddite.pro/>
4. SATO, Mia. Nick Clegg says asking artists for use permission would 'kill' the AI industry. En: *The Verge* [en línea]. 26 de mayo de 2025 [consulta: 3 de noviembre de 2025]. Disponible en: <https://www.theverge.com/news/674366/nick-clegg-uk-ai-artists-policy-letter>
5. SATO SCHAUL, Kevin Schaul; CHEN, Szu Yu; TIKU, Nitasha. Inside the secret list of websites that make AI like ChatGPT sound smart. En: *The Washington Post* [en línea]. 19 de abril de 2023 [consulta: 5 de mayo de 2023]. Disponible en: <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>
6. SATO MILMO, Dan. 'Impossible' to create AI tools like ChatGPT without copyrighted material, OpenAI says. En: *The Guardian* [en línea]. 8 de enero de 2024 [consulta: 8 de enero de 2024]. Disponible en: <https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material-openai>
7. MARCUS, Gary; SOUTHEN, Reid. Generative AI Has a Visual Plagiarism Problem Experiments with Midjourney and DALL-E 3 show a copyright minefield. En: *IEEE SPECTRUM* [en línea]. 6 de enero de 2024 [consulta: 7 de enero de 2024]. Disponible en: <https://spectrum.ieee.org/midjourney-copyright>
8. THIEL, David. Investigation Finds AI Image Generation Models Trained on Child Abuse. En: *STANFORD The Cyber Policy Center* [en línea]. 20 de diciembre de 2023 [consulta: 10 de enero de 2024]. Disponible en: <https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse>
9. <https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse>
10. EU Artificial Intelligence Act [fecha de consulta: 13 de junio de 2024]. Disponible en: <https://artificialintelligenceact.eu/es/>
11. CISAC. *Estudio de PMP Strategy/CISAC sobre el impacto económico de la IA en las industrias musical y audiovisual* [en línea]. Noviembre de 2024. Disponible en: <https://www.cisac.org/es/servicios/informes-y-estudios/estudio-de-ia-cisacpmp-strategy>

Dra. Chele Esteve Sendra, Ángela Muñoz Iglesias y Dr. Rafel Arnal Rodrigo.
Inteligencia Artificial, accesibilidad y diseño inclusivo en el metaverso: Hacia una
justicia tecnológica en la era 5IR

Dra. Chele Esteve Sendra

Universitat Politècnica de València

Ángela Muñoz Iglesias

Universitat Politècnica de València

Dr. Rafel Arnal Rodrigo

Universidad de Zaragoza

[ACCEDE A LA GRABACIÓN](#)

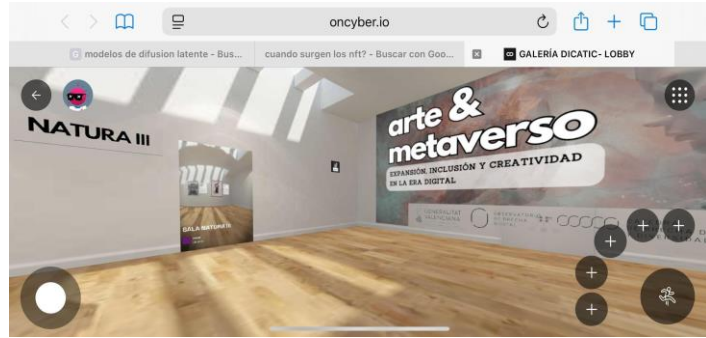
RESUMEN

Esta comunicación presenta una investigación aplicada desarrollada en el marco de la **Cátedra de Brecha Digital y Discapacidad (DicaTIC)** de la **Universitat Politècnica de València**, centrada en la creación y análisis de **exposiciones en el metaverso** (Figuras. 1 -2) con el fin de evaluar su **accesibilidad** y el papel de la **inteligencia artificial (IA)** y el **diseño inclusive** en la construcción de **entornos digitales accesibles**. El estudio examina dos plataformas: **Spatial.io** y **Oncyber.io**, a partir de los **principios de diseño universal** y las **pautas internacionales de accesibilidad (WCAG 2.1)**, mediante **pruebas de usabilidad** con personas con discapacidad visual, auditiva y motora, complementadas con **entrevistas a expertos** en accesibilidad y diseño ético. Los resultados evidencian **carencias significativas** en la navegación por teclado, la lectura por pantalla, el subtítulo en tiempo real y las descripciones de audio, lo que pone de manifiesto una **exclusión estructural** en los entornos inmersivos actuales. Se formulan **recomendaciones** para integrar la accesibilidad desde las fases iniciales de diseño a través de **IA adaptativa**, **procesos participativos** y **co-diseño con comunidades afectadas**. El trabajo contribuye a los **Objetivos de Desarrollo Sostenible (ODS)**; especialmente el **ODS 9 (Industria, innovación e infraestructura)**, el **ODS 10 (Reducción de las desigualdades)** y el **ODS 11 (Ciudades y comunidades sostenibles)**, al promover un modelo de metaverso equitativo, inclusivo y centrado en la diversidad.

PALABRAS CLAVE: accesibilidad digital; diseño universal; inteligencia artificial; metaverso; discapacidad; justicia tecnológica.



CONVOCATORIA DE EXPOSICIÓN VIRTUAL
CREACIÓN DE DOS EXPOSICIONES EN EL METAVERSO



Figuras 1 y 2. Cartel y vista de la galería virtual “Arte & Metaverso: Expansión, inclusión y creatividad en la era digital” (2024), promovida por la Cátedra DicaTIC de la Universitat Politècnica de València con el apoyo de la Generalitat Valenciana. Las salas, alojadas en Oncyber.io, se conciben como espacios experimentales de accesibilidad e inclusión en el metaverso. Fuente: Elaboración propia, 2024.

1. INTRODUCCIÓN

La denominada **Quinta Revolución Industrial (5IR)** se caracteriza por la convergencia entre la **inteligencia artificial**, la **robótica colaborativa** y una **orientación humanista de la tecnología**. A diferencia de la Cuarta Revolución Industrial, centrada en la automatización, la **5IR** busca reconciliar la innovación con la sostenibilidad y la inclusión social. En este contexto, el **metaverso**, entendido como un entorno digital tridimensional, interactivo y persistente, ofrece nuevas oportunidades para el aprendizaje, la creación y la socialización, pero también plantea **desafíos éticos y de accesibilidad** que es necesario abordar.

El acceso equitativo a estos entornos virtuales no está garantizado. Las **personas con discapacidad** siguen encontrando **barreras técnicas, simbólicas y de representación** que dificultan su plena participación en la vida digital. La accesibilidad, lejos de ser un aspecto técnico secundario, constituye un **derecho fundamental**, reconocido en la **Convención sobre los Derechos de las Personas con Discapacidad** (ONU, 2006) y en las políticas europeas sobre **brecha y derechos digitales**. Este trabajo se enmarca en las investigaciones desarrolladas por la **Cátedra DicaTIC de la UPV**, cuyo objetivo es promover el **diseño inclusivo** y la **transformación ética del entorno digital**. Desde 2021, la Cátedra impulsa proyectos, publicaciones y exposiciones que exploran la intersección entre **arte, música, educación, tecnología y discapacidad**, generando espacios donde la **creación**, la **investigación aplicada** y la **participación social** se conciben como instrumentos de **inclusión y accesibilidad universal**.

El presente estudio **analiza la accesibilidad del metaverso** a través de la evaluación de dos plataformas: *Spatial.io* y *Oncyber.io*, empleadas como soporte para **siete salas expositivas accesibles desarrolladas por la Cátedra DicaTIC**. Mediante un **enfoque cualitativo de investigación-acción**, se identifican **barreras visuales, auditivas, motrices y cognitivas**, y se experimenta con una **IA generativa** para la **audiodescripción automática**, destinada a mejorar la orientación de personas con discapacidad visual. Las salas fueron creadas específicamente para **evaluar su grado de accesibilidad** con grupos de personas con discapacidad y, al mismo tiempo, para dar a conocer las **obras de artistas con y sin discapacidad**, fomentando la inclusión y la participación en entornos digitales compartidos.

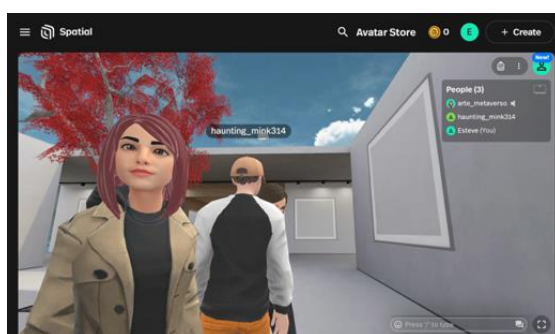
El artículo tiene como propósito reflexionar sobre la necesidad de **integrar la accesibilidad desde el diseño inicial** y redefinir el papel de la **inteligencia artificial** como **mediadora sensorial y herramienta de justicia tecnológica**. Los resultados contribuyen a delinear un modelo de **metaverso inclusivo, ético y participativo**, alineado con los **Objetivos de Desarrollo Sostenible** (ODS 9, 10 y 11), y con la misión de la **Universitat Politècnica de València** de avanzar hacia una **transformación digital justa y sostenible**.

2. MARCO TEÓRICO: ACCESIBILIDAD, INTELIGENCIA ARTIFICIAL Y DERECHOS DIGITALES

La **accesibilidad universal** implica que toda persona, con independencia de sus capacidades, pueda **percibir, comprender, navegar e interactuar** con los sistemas digitales (ISO 9241-210, 2019). Este principio se vincula con el **diseño universal**, orientado a crear productos y entornos utilizables por el mayor número posible de personas sin necesidad de adaptaciones posteriores. La **Convención de las Naciones Unidas sobre los Derechos de las Personas con Discapacidad** (ONU, 2006) reconoce el **acceso a la información y la comunicación** como un derecho fundamental, estableciendo las bases para el desarrollo de entornos digitales inclusivos. En el contexto europeo, la **European Accessibility Act** (2019) y el **AI Act** (2021) obligan a las instituciones públicas y privadas a garantizar que las tecnologías sean **seguras, transparentes y no discriminatorias**, reforzando así la dimensión ética y jurídica de la accesibilidad digital. La **ética de la inteligencia artificial (IA)** ha sido analizada por Floridi (2019), quien subraya que los principios de **justicia, beneficencia y no maleficencia** deben traducirse en **prácticas concretas de diseño y gobernanza algorítmica**. En el ámbito español, Mañas-Viniegra, Rodríguez-Fernández, Herrero-de-la-Fuente y Veloso (2023) examina la **responsabilidad social del comunicador** y su papel en la integración de tecnologías accesibles, destacando la necesidad de una **alfabetización digital ética e inclusiva** que acompañe la innovación tecnológica.

2.1 Metaverso y justicia tecnológica

El **metaverso** se configura como un nuevo espacio de interacción en el que confluyen la **arquitectura digital**, la **simulación social** y la **economía simbólica**. Su diseño plantea inevitables **cuestiones éticas y políticas**: ¿Quién puede habitar estos entornos? ¿Qué cuerpos y subjetividades resultan representables como avatar (Figuras 3-4)? ¿Qué tecnologías median y condicionan la experiencia inmersiva?



Figuras 3 y 4. Interacción de usuarios en una exposición virtual dentro de la plataforma Spatial.io. La experiencia evidencia las limitaciones actuales de personalización y accesibilidad en entornos inmersivos, donde la representación corporal continúa siendo homogénea y excluyente para personas con discapacidad. Imagen: Chele Esteve, 2024.

Desde la perspectiva de la **justicia tecnológica** (Costanza-Chock, 2020), el desarrollo digital debe orientarse a **corregir desigualdades estructurales**, reconociendo la **discapacidad** no como una mera condición biomédica, sino como una **categoría política** que interpela la distribución del poder, la accesibilidad y la representación dentro de los sistemas sociotécnicos. Este enfoque reclama una **tecnología inclusiva y participativa**, que incorpore la **diversidad humana en la propia concepción y gobernanza del metaverso**.

2.2 Inteligencia artificial y diseño inclusivo

La **IA adaptativa** permite personalizar interfaces según las necesidades del usuario: lectores de pantalla contextuales, ajustes automáticos de contraste, reconocimiento de voz, etc. Sin embargo, su eficacia depende de que los algoritmos se entrenen con **datos inclusivos** (Figura. 5) y de que los procesos de diseño incorporen la voz de las personas afectadas ⁷.



Figura. 5. Ecosistema conceptual de la investigación: IA – accesibilidad – metaverso – derechos humanos. Imagen generada con inteligencia artificial. Fuente: Elaboración propia a partir de ISO 9241-210 (2019), ONU (2006), Floridi (2019) y Mañas-Viniegra, Rodríguez-Fernández, Herrero-de-la-Fuente y Veloso (2023).

⁷ El proyecto se enmarca en las líneas de investigación de la Cátedra DicaTIC, centradas en la inclusión digital y el diseño universal.

3. METODOLOGÍA

3.1 Enfoque general

El estudio se desarrolló mediante un **método cualitativo de investigación-acción**, orientado a la **mejora práctica de la accesibilidad digital** en entornos inmersivos. En una primera fase, se realizó una **revisión exploratoria de distintas plataformas del metaverso**, con el objetivo de identificar aquellas que mostraran **mayor potencial de accesibilidad** para personas con **diversas discapacidades**: visual, auditiva y motora. A partir de este análisis inicial, se seleccionaron **Spatial.io** y **Oncyber.io** por su relevancia en el ámbito **artístico y educativo**, así como por las herramientas de interacción y personalización que ofrecen. Posteriormente, se llevaron a cabo pruebas de accesibilidad siguiendo los principios de las Web Content Accessibility Guidelines (WCAG 2.1), con la participación de nueve usuarios con discapacidad: tres visuales, tres auditivas y tres motoras, junto con cinco expertos en accesibilidad digital y diseño UX (Figura.



Figura. 6. Fases del estudio: selección → evaluación → codiseño → recomendaciones. Diagrama generado con inteligencia artificial (IA). Fuente: elaboración propia, 2025.

Además, se realizaron reuniones participativas (presenciales y en línea) con representantes de distintos colectivos de personas con discapacidad, coordinadas a través del **Comité Español de Representantes de Personas con Discapacidad de la Comunitat Valenciana (CERMI CV)** y la **Fundación ONCE**, quienes validaron los procedimientos y resultados de las pruebas. Agradecemos especialmente su colaboración desinteresada y su compromiso con la mejora de la accesibilidad digital. Este enfoque colaborativo permitió **evaluar comparativamente la experiencia de uso en ambas plataformas**, así como **identificar barreras y oportunidades de mejora**, consolidando un **modelo de investigación** centrado en el usuario y orientado a la construcción de un **metaverso verdaderamente inclusivo**.

3.2 Participantes y herramientas

En el estudio participaron nueve **usuarios con discapacidad**, tres con discapacidad visual, tres con auditiva y tres con motora, así como cinco **expertos en accesibilidad digital y diseño de experiencia de usuario (UX)**. Las pruebas se realizaron entre **febrero y mayo de 2024**. Para el

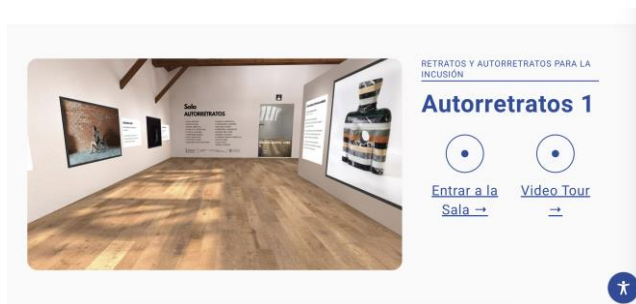
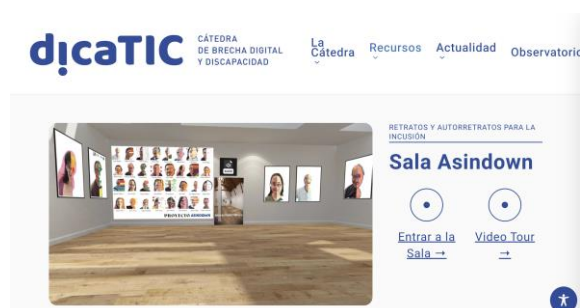
desarrollo de las sesiones se emplearon diversas **herramientas de apoyo y registro** audiovisual y el análisis de usabilidad.

3.3 Criterios de evaluación

La evaluación se basó en los **cuatro principios fundamentales de las Pautas de Accesibilidad para el Contenido Web (WCAG 2.1)**: **perceptible**, que establece que la información debe poder percibirse a través de cualquiera de los sentidos; **operable**, que exige que la interfaz pueda utilizarse mediante distintos métodos de interacción como el teclado, el lector de pantalla o el control por voz; **comprensible**, que requiere que el contenido sea claro, coherente y predecible; y **robusto**, que garantiza la compatibilidad del entorno con diferentes tecnologías y ayudas técnicas. Estos criterios se aplicaron al análisis de las salas virtuales desarrolladas (Figuras. 7 y 8), donde se evaluó su grado de accesibilidad y la experiencia de interacción de los usuarios en

contextos

inmersivos.



Figuras. 7 y 8. Sala Autorretratos 1 (izquierda) y Sala Asindown (derecha). La participación del grupo Asindown Envejecimiento Activo en el Proyecto Retratos y Autorretratos para la Inclusión fue tan significativa que motivó la creación de una sala virtual específica titulada “Sala Asindown. Retratos y Autorretratos para la Diversidad”, donde se reúnen sus autorretratos y testimonios visuales, integrados en un recorrido artístico en proceso de mejora hacia una accesibilidad plena.

3.4 Análisis de datos

Los datos se analizaron de forma cualitativa a partir de los informes de observación y las hojas de validación completadas por los participantes y expertos (Tabla 1). La información se organizó en tres categorías principales: barreras técnicas, barreras simbólicas y barreras de

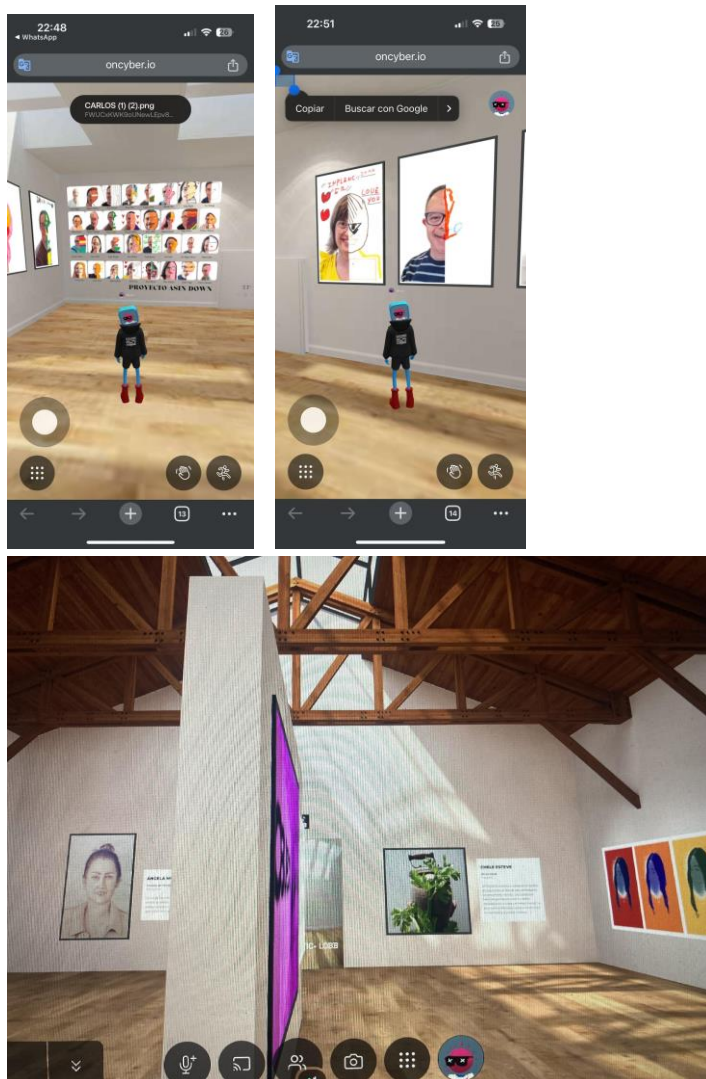
representación, identificando patrones comunes en la experiencia de uso y accesibilidad de ambas plataformas. Las entrevistas y registros de interacción se revisaron conjuntamente para asegurar la coherencia y fiabilidad de los resultados.

Dimensión	Indicadores evaluados	Spatial.io	Oncyber.io
Visual	Contraste / lector de pantalla / texto alternativo	 Parcial	 Parcial
Auditiva	Subtitulado / alertas visuales / control de volumen	 Insuficiente	 Aceptable
Motriz	Navegación por teclado / control por voz / interfaces	 Limitado	 Mejorado
Cognitiva	Claridad de navegación / tutoriales /	 Aceptable	 Aceptable

Tabla 1. Resultados comparativos de accesibilidad por tipo de discapacidad que muestra los indicadores evaluados en las plataformas Spatial.io y Oncyber.io según cuatro dimensiones (visual, auditiva, motriz y cognitiva). Fuente: Elaboración propia, 2025.

3.5 Limitaciones

El **estudio** se centró en dos plataformas concretas (Figuras. 9, 10 y 11) y en un número reducido de usuarios; por tanto, los resultados no pretenden generalizarse sino identificar patrones de exclusión replicables en otros entornos.



Figuras 9, 10 y 11. Vista de la aplicación móvil Oncyber.io de la Sala Asindown (izquierda, Figuras 9 y 10) y Sala Autorretratos 2 desde ordenador (derecha). Oncyber.io, plataforma gratuita para crear espacios expositivos en el metaverso, se utilizó en este proyecto para explorar la accesibilidad digital. En esta experiencia participaron personas con y sin discapacidad, junto a artistas de reconocido prestigio que defienden que la inclusión es un derecho y que todas las personas somos iguales. Fuente: Chele Esteve, 2024.

4. RESULTADOS Y ANÁLISIS

Las pruebas revelaron que ninguna de las plataformas cumple plenamente con los estándares WCAG 2.1.

En *Spatial.io* se observó una interfaz más intuitiva, pero con fallos en descripciones ALT y subtítulo.

Oncyber.io, en cambio, presenta mejor compatibilidad con lectores de pantalla y permite una navegación más lineal, aunque requiere registro y presenta sobrecarga visual.



Tabla 2. Mapa comparativo de barreras detectadas en ambas plataformas que muestra las barreras técnicas, simbólicas y de representación identificadas durante las pruebas de accesibilidad en *Spatial.io* y *Oncyber.io*. Fuente: Elaboración propia, 2025.

Las principales limitaciones detectadas en las plataformas analizadas fueron la **falta de integración con tecnologías de asistencia**, la **navegación compleja sin uso de ratón**, la **ausencia de subtítulos y descripciones de audio**, y la **escasa posibilidad de personalizar avatares que representen la diversidad funcional**. Estas deficiencias generan una forma de **exclusión simbólica**, en la que el usuario con discapacidad es percibido como un “visitante tolerado” más que como un agente activo **dentro del entorno virtual**.

4.1 Percepciones de los usuarios

Los participantes destacaron la frustración de no poder desplazarse libremente o comprender el entorno sin apoyo visual. Una participante con discapacidad auditiva señaló: *“El metaverso tiene música, luces, interacción... pero si no oigo ni puedo leer, solo veo avatares moviéndose sin sentido.”* Del mismo modo, un usuario con discapacidad visual comentó: *“Puedo acceder a muchas webs con mi lector de pantalla, pero aquí no hay voces ni texto alternativo. El metaverso es un lugar silencioso y vacío para mí.”* Estas percepciones reflejan la brecha sensorial y cognitiva

existente en los entornos inmersivos actuales, así como la necesidad de **incorporar accesibilidad desde el diseño inicial** para garantizar una experiencia equitativa para todos los usuarios.

4.2 Resultados de accesibilidad

Las **pruebas de accesibilidad** realizadas en las plataformas *Spatial.io* y *Oncyber.io* confirman que ambas presentan **carencias estructurales** en el cumplimiento de las *Pautas de Accesibilidad para el Contenido Web* (WCAG 2.1). Ninguna de las dos alcanza el **nivel AA** exigido por la normativa europea de accesibilidad digital. La evaluación permitió identificar un conjunto de **barreras funcionales y simbólicas**, agrupadas en cuatro dimensiones: **visual, auditiva, motriz y cognitiva**.

4.3 Barreras visuales

Los **lectores de pantalla NVDA 2023 y JAWS 2024** detectaron deficiencias graves en el **etiquetado de elementos gráficos** y en la **jerarquía semántica** del código HTML subyacente. En *Spatial.io*, el **contraste entre texto y fondo** no supera la relación 4.5:1 recomendada por las WCAG para textos normales. Los usuarios con discapacidad visual reportaron una **sensación de desorientación espacial** y la **ausencia de puntos de referencia sonoros o vibratorios** que facilitaran la orientación y la navegación.

4.4 Barreras auditivas

Las dos plataformas analizadas carecen de **subtitulado automático fiable y de alertas visuales** que sustituyan las señales sonoras. En *Oncyber.io* se comprobó que la funcionalidad de chat visual podía paliar parcialmente esta limitación, aunque sin sincronización temporal con el audio. La **inaccesibilidad auditiva** afecta también a la dimensión educativa, al impedir la comprensión plena de conferencias o exposiciones con audio especializado.

4.5 Barreras motrices

Las pruebas con usuarios que empleaban dispositivos alternativos de entrada; trackball, control por voz o seguimiento ocular, revelaron que **ninguna interfaz permite la navegación completa sin ratón**. Los atajos de teclado son inconsistentes y no existe un modo “accesible” que simplifique la interacción.

4.6 Barreras cognitivas

La complejidad visual y la abundancia de estímulos dinámicos dificultan la comprensión de la interfaz, especialmente en usuarios con trastornos de atención o dificultades de lectura. No se ofrece un modo simplificado ni glosarios visuales para interpretar iconografía específica.

4.7 Percepciones de los usuarios

Los testimonios recogidos destacan el contraste entre la expectativa de inclusión y la realidad de la exclusión digital (Tabla 3). Una participante con discapacidad auditiva afirmó: *“El metaverso promete conexión, pero si no puedo leer ni escuchar, no estoy dentro; solo mi avatar se mueve.”* Los expertos en accesibilidad coincidieron en que la falta de estándares específicos para entornos 3D explica la fragmentación actual de soluciones.

Dimensión	Principales barreras	Impacto en la experiencia
Visual	Contraste insuficiente, ausencia de descripciones ALT	Desorientación espacial
Auditiva	Ausencia de subtítulos y alertas visuales	Pérdida de información verbal
Motriz	Navegación por teclado incompleta	Imposibilidad de uso autónomo
Cognitiva	Sobrecarga de información visual	Fatiga y abandonos prematuros

Tabla 3. Síntesis de barreras de accesibilidad detectadas. Presenta las principales barreras visuales, auditivas, motrices y cognitivas, junto con su impacto en la experiencia del usuario. Fuente: Elaboración propia, 2025.

4.8 Síntesis analítica

Los resultados evidencian que la accesibilidad sigue considerándose un complemento técnico y no un criterio estructural. La evaluación de Spatial.io y Oncyber.io pone de manifiesto la necesidad de **redefinir la accesibilidad como principio de diseño ético y de justicia tecnológica**.

5. Discusión

Los resultados confirman que la accesibilidad no es aún prioridad en el desarrollo de mundos virtuales. La **brecha digital** se reconfigura en el **plano sensorial y cognitivo**.

5.1 IA y participación

La inteligencia artificial **adaptativa** ofrece un gran potencial para **personalizar la experiencia de usuario**, detectando automáticamente las preferencias, necesidades y modos de interacción de cada participante. En contextos inclusivos, puede ajustar el nivel de detalle visual, la velocidad de lectura o la complejidad del lenguaje, favoreciendo una **accesibilidad dinámica y contextual**. En el marco del proyecto se desarrollaron **siete salas virtuales accesibles**, en lo posible, dentro de la plataforma **Oncyber.io**, concebidas como **espacios expositivos experimentales** para analizar la accesibilidad del metaverso desde una perspectiva **educativa, artística y social**. Estas salas con temática de **autorretratos, naturaleza y mixturas** pueden **visitarse libremente en Oncyber⁸**, y constituyen un laboratorio digital para reflexionar sobre la representación, la interacción y la participación de personas con discapacidad en entornos inmersivos. El proyecto fue coordinado por la **Cátedra DicaTIC** de la **UPV**, con la participación de entidades como la **Fundación ONCE** y **CERMI CV**, que colaboraron en las fases de **validación y evaluación**. En las pruebas participaron **usuarios con discapacidad visual, auditiva y motriz**, junto con **expertos en accesibilidad y diseño inclusivo**, quienes aportaron **retroalimentación directa** sobre la **usabilidad, la comprensión espacial y la percepción de las salas virtuales**. Durante las sesiones se implementó de forma experimental una **IA generativa para audiodescripción automática**, capaz de **narrar en tiempo real las obras expuestas y los elementos visuales y espaciales del entorno** a personas con **discapacidad visual**. Con el fin de ofrecer una solución específica para personas con discapacidad visual, se **diseñó e integró expresamente un video tour accesible** que guiaba al visitante a través de las salas virtuales, proporcionando **descripciones dinámicas de las obras, los recorridos y los puntos de referencia**. Esta herramienta se desarrolló para **mejorar la orientación espacial y la experiencia sensorial** dentro del metaverso. El recorrido continúa disponible a través del **enlace alojado en la web de la Cátedra DicaTIC de la UPV**. El **modelo piloto** demostró el **potencial de la inteligencia artificial como mediadora sensorial y herramienta de inclusión cultural**, al facilitar la **orientación, la comprensión espacial y la participación active** de los usuarios en entornos digitales. La experiencia evidenció, además, la necesidad de **incorporar la accesibilidad desde las fases iniciales del diseño y de concebir la audiodescripción no como un añadido, sino como un elemento estructural del discurso expositivo**. No obstante, esta misma capacidad tecnológica plantea **dilemas éticos** relacionados con la **privacidad de los datos**

⁸ <https://dicatic.webs.upv.es/artemetaverso/> (consulta: 30 de septiembre de 2025).

sensoriales y biométricos utilizados en los modelos predictivos, así como con la **responsabilidad en la interpretación automatizada de la experiencia estética y emocional** (Floridi, 2019).

5.2 Accesibilidad como valor añadido

La **accesibilidad**, lejos de constituir una mera obligación legal, representa un **valor estratégico y una ventaja competitiva** en el desarrollo de entornos digitales. Las **plataformas inclusivas** no solo amplían su base de usuarios, sino que también **refuerzan su legitimidad social** al responder a los principios de equidad y participación universal. Además, la **incorporación de personas con discapacidad en los procesos de co-diseño** contribuye a mejorar la **calidad de la interfaz**, optimizar la **experiencia de usuario** y **reducir los costes de mantenimiento y adaptación** a largo plazo.

6. SOLUCIONES

Los hallazgos obtenidos invitan a reflexionar sobre la **tensión entre innovación tecnológica y equidad digital**, evidenciando que la **accesibilidad** no depende únicamente del **cumplimiento de estándares técnicos**, sino de la **voluntad ética y cultural** de integrar la diversidad en todas las fases del diseño. Este enfoque subraya que la verdadera innovación no reside solo en el avance tecnológico, sino en la capacidad de crear **entornos digitales inclusivos, sostenibles y socialmente responsables**.

6.1 El metaverso como nuevo espacio de exclusión o inclusión

El **metaverso** tiende a **amplificar las brechas preexistentes** al trasladarlas al plano virtual. Las **dificultades de navegación, comunicación y representación** reproducen en el entorno digital los mismos **obstáculos que persisten en el espacio físico**, perpetuando desigualdades estructurales. Como señala Costanza-Chock (2020), las **tecnologías inmersivas no son neutrales**, sino que reflejan los **sesgos de quienes las diseñan y programan**. En este contexto, la **justicia tecnológica** implica **democratizar la producción de los entornos virtuales** y garantizar la **participación activa de los grupos históricamente marginados** en su concepción y desarrollo.

6.2 La IA adaptativa como oportunidad inclusiva

La **inteligencia artificial** ofrece **herramientas prometedoras** para la mejora de la accesibilidad, como el **reconocimiento de voz**, la **generación automática de subtítulos** o los **ajustes de contraste**

mediante visión por computador. Sin embargo, estas soluciones deben evaluarse desde una perspectiva ética, considerando especialmente la **privacidad y el tratamiento responsable de los datos personales**. La integración de **algoritmos adaptativos** capaces de aprender de las **preferencias y necesidades del usuario** representa un **avance decisivo hacia entornos digitales verdaderamente centrados en el ser humano**, donde la tecnología actúe como mediadora para la inclusión y no como fuente de nuevas desigualdades.

6.3 De la responsabilidad técnica a la responsabilidad social

La **accesibilidad** no puede entenderse únicamente como el **cumplimiento de la legislación vigente**, sino como una **responsabilidad social inherente al diseño**. Como señalan; Mañas-Viniegra, Rodríguez-Fernández, Herrero-de-la-Fuente y Mañas-Viniegra, Rodríguez-Fernández, Herrero-de-la-Fuente y Veloso (2023), la accesibilidad constituye un **valor añadido** que enriquece la **experiencia de todos los usuarios**, no solo de aquellos con discapacidad. En el **contexto universitario**, la **inclusión digital** refuerza la **misión pública** de las instituciones de educación superior y su **compromiso con los Objetivos de Desarrollo Sostenible (ODS 9, 10 y 11)**, promoviendo una transformación tecnológica **ética, equitativa y sostenible**.

6.4 Limitaciones y perspectivas

Entre las **principales limitaciones** del estudio se encuentran la **muestra reducida de participantes** y el **análisis centrado en solo dos plataformas**. No obstante, los resultados proporcionan una **base empírica sólida** para avanzar en la definición de **estándares de accesibilidad específicos para entornos XR y metaversos**. Las **futuras líneas de investigación** deberán profundizar en el **potencial de la IA generative** para el diseño de **interfaces adaptativas** y en la **inclusión sensorial** a través de **tecnologías hápticas**⁹.

⁹ Las tecnologías hápticas son sistemas que reproducen sensaciones táctiles o de presión mediante vibraciones, fuerzas o movimientos, permitiendo al usuario “sentir” objetos o texturas en entornos digitales o virtuales. En este contexto, se utilizan para mejorar la accesibilidad sensorial, facilitando que personas con discapacidad visual, por ejemplo, perciban información espacial o artística a través del tacto o la retroalimentación física.



Tabla 4. Síntesis visual de los principios y recomendaciones finales. Recoge los principios clave de accesibilidad, diseño inclusivo y justicia tecnológica, junto con las recomendaciones para un metaverso ético y participativo derivadas del estudio. Fuente: Elaboración propia, 2025.

7. RECOMENDACIONES PARA UN METAVERSO INCLUSIVO

Los resultados obtenidos permiten formular una serie de **recomendaciones técnicas, pedagógicas y éticas** destinadas a orientar el desarrollo de entornos inmersivos verdaderamente accesibles.

7.1 Integrar la accesibilidad desde la fase inicial

La accesibilidad no debe añadirse al final del proceso de diseño, sino formar parte de la concepción inicial de cualquier producto o entorno digital. Incluirla desde el principio reduce costes y garantiza coherencia funcional y estética.

7.2 Participación activa y co-diseño

La **participación de personas con discapacidad en la planificación, desarrollo y validación de los entornos inmersivos** resulta esencial. Su experiencia directa permite identificar barreras invisibles para los diseñadores convencionales. La metodología del **co-diseño** empleada en la Cátedra DicaTIC ha demostrado que los procesos colaborativos mejoran la calidad y pertinencia de las soluciones desarrolladas.

7.3 Aplicación de inteligencia artificial adaptativa

Se recomienda desarrollar algoritmos capaces de reconocer patrones de interacción y ajustar dinámicamente la interfaz a las necesidades del usuario. Ejemplos: ajuste automático de contraste, ampliación de texto, subtitulado instantáneo o navegación mediante comandos de voz. Sin embargo, la IA debe programarse bajo criterios de **transparencia, control humano y protección de datos personales**, conforme al Reglamento Europeo de Inteligencia Artificial (AI Act, 2021).

7.4 Estandarización de la accesibilidad inmersiva

Urge crear estándares internacionales que complementen las **WCAG 2.1** y se adapten a las especificidades de los entornos tridimensionales. Dichos estándares deberían definir niveles de accesibilidad en función de la compatibilidad con dispositivos hápticos, lectores de pantalla 3D y sistemas de realidad aumentada.

7.5 Formación y sensibilización

La inclusión tecnológica requiere formación transversal en accesibilidad digital para diseñadores, ingenieros, docentes y responsables de innovación. Las universidades deben incorporar estos contenidos en los programas de diseño y arquitectura, fomentando la responsabilidad social y la ética del diseño. Los resultados obtenidos permiten formular una serie de **recomendaciones técnicas, pedagógicas y éticas** destinadas a orientar el desarrollo de entornos inmersivos verdaderamente accesibles. Estas recomendaciones se orientan a construir una **ética del diseño virtual**, donde la equidad se traduzca en decisiones tecnológicas concretas.

Categoría	Recomendación clave	Objetivo
 Diseño	Integrar accesibilidad desde el inicio	Prevenir exclusión
 Participación	Co-diseñar con personas con discapacidad	Legitimar resultados
 IA	Implementar IA adaptativa y transparente	Personalizar experiencias
 Estándares	Desarrollar pautas específicas para XR	Homogeneizar criterios
 Formación	Capacitar en accesibilidad y ética digital	Fomentar responsabilidad

Tabla 5. Recomendaciones técnicas y éticas para entornos inmersivos inclusivos. Fuente: Elaboración propia, 2025.

8. CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN

El **metaverso** se configura como un **laboratorio privilegiado** para analizar la interacción entre **tecnología y derechos humanos**. Su desarrollo no puede desvincularse de los **principios de equidad, participación y sostenibilidad** que inspiran la **Quinta Revolución Industrial (5IR)**. La investigación evidencia que la **accesibilidad** continúa siendo un **reto pendiente**, pero también una **oportunidad para reimaginar la tecnología como un espacio de justicia social**. Un **metaverso inclusivo** no solo elimina barreras, sino que **amplía las posibilidades de participación cultural, educativa y creativa**, promoviendo la diversidad y la equidad digital. En este sentido, el **acceso al metaverso** debería reconocerse como un **nuevo derecho digital**, y garantizar su accesibilidad implica no solo **cumplir con la normativa**, sino avanzar hacia una **justicia tecnológica y una democracia cultural efectivas**.

Las futuras líneas de investigación se orientarán a:

- La **integración de inteligencia artificial adaptativa** en plataformas educativas inmersivas.
- La **evaluación del impacto emocional** de la inclusión virtual en diferentes perfiles de usuario.
- El **desarrollo de guías de buenas prácticas** para el diseño y la gestión de **metaversos accesibles**.

8.1 Síntesis final

Las **plataformas inmersivas actuales** aún no cumplen con los **estándares internacionales de accesibilidad**, lo que evidencia la necesidad de una **transformación profunda en el diseño digital**. En primer lugar, la **accesibilidad** debe asumirse como una **dimensión ética**, además de técnica, integrada desde las fases iniciales del desarrollo. En segundo lugar, la **inteligencia artificial adaptative** representa un **potencial transformador**, siempre que se fundamente en **principios de transparencia, participación y co-creación**. Asimismo, las **universidades, cátedras especializadas y asociaciones del ámbito de la discapacidad**, como el **CERMI Comunidad Valenciana** y la **Fundación ONCE**, están llamadas a **liderar la investigación aplicada** y a actuar como **mediadoras** entre la **innovación tecnológica y la justicia social**. Finalmente, la **colaboración entre administración pública, comunidad académica y sociedad civil** resulta esencial para avanzar hacia una **verdadera equidad tecnológica** y garantizar que los **entornos digitales del future** sean **inclusivos, sostenibles y humanistas**.



Figuras 12 y 13. Captura de pantalla de la página web de la Cátedra DicaTIC (arriba) y vista interior de una de las salas virtuales del metaverso (abajo). La web presenta el proyecto, que reúne más de 120 obras artísticas distribuidas en siete salas virtuales, y pone en valor el papel del arte como herramienta de comunicación, sensibilización y conciencia sobre la inclusión y la diversidad. En la sala virtual inferior se aprecia la interacción mediante avatars en un entorno inmersivo en proceso de mejora hacia una accesibilidad plena. Fuente: Ángela Muñoz, 2024.

8.2 Líneas de investigación futuras

Las futuras líneas de investigación se orientan hacia el **desarrollo de interfaces multisensoriales** que integren **estímulos visuales, sonoros y hápticos adaptables**, favoreciendo una experiencia inmersiva más inclusiva. Asimismo, se propone la **evaluación del impacto psicológico y social** de la inclusión virtual, con el fin de comprender cómo estas tecnologías influyen en la percepción, la identidad y la participación de los usuarios. Otra línea prioritaria es el **diseño de metaversos educativos** centrados en la **accesibilidad cognitiva** y el **aprendizaje experiencial**, que permitan nuevas formas de enseñanza participativa (MURILLO I RIBES et al., 2023). Finalmente, se plantea la **elaboración de guías internacionales de accesibilidad XR**, impulsadas desde la **Unión Europea** y organismos como la **UNESCO**, para establecer marcos de referencia comunes y garantizar una transformación digital ética y universal.

BIBLIOGRAFÍA

COSTANZA-CHOCK, S. *Design justice: community-led practices to build the worlds we need* [en línea]. Cambridge (MA): MIT Press, 2020. ISBN 978-0-262-04345-8. Disponible en: <https://mitpress.mit.edu/9780262043458/design-justice/> [Consulta: 30 septiembre 2025].

EUROPEAN UNION. *Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services (European Accessibility Act)* [en línea]. Bruselas: Unión Europea, 2019. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0882> [Consulta: 30 septiembre 2025].

EUROPEAN COMMISSION. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)* [en línea]. Bruselas: Comisión Europea, 2021. COM (2021) 206 final. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206> [Consulta: 30 septiembre 2025].

FLORIDI, L. Translating principles into practices of digital ethics: five risks of being unethical. *Philosophy & Technology* [en línea]. 2019, vol. 32, n.º 2, pp. 185–193. ISSN 2210-5433. DOI: <https://doi.org/10.1007/s13347-019-00354-x> [Consulta: 30 septiembre 2025].

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems* [en línea]. Ginebra: ISO, 2019. Disponible en: <https://www.iso.org/standard/77520.html> [Consulta: 30 septiembre 2025].

MAÑAS-VINIEGRA, L. et al. Nuevas tecnologías aplicadas a la comunicación, educación y empleabilidad para la inclusión de las personas con discapacidad en la sociedad digital. *Revista ICONO 14* [en línea]. 2023, vol. 21, n.º 2, pp. e2047. ISSN 1697-8293. DOI: <https://doi.org/10.7195/ri14.v21i2.2047> [Consulta: 30 septiembre 2025].

MURILLO I RIBES, A.; GONZÁLEZ ZAMORA, A.; TEJADA GIMÉNEZ, J. Percepción del profesorado sobre el uso de Soundcool como herramienta tecnológica en la formación inicial del profesorado. *Revista Educare* [en línea]. 2023, vol. 26, n.º 1. ISSN 1409-4258. Disponible en: <https://www.revistaeducare.ucr.ac.cr/> [Consulta: 30 septiembre 2025].

NACIONES UNIDAS. *Convención sobre los Derechos de las Personas con Discapacidad* [en línea]. Nueva York: ONU, 2006. Disponible en: <https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html> [Consulta: 30 septiembre 2025].

UNESCO. *Ethical framework for AI in education* [en línea]. París: UNESCO, 2023. Disponible en: <https://unesdoc.unesco.org/ark:/48223/pf0000386135> [Consulta: 30 septiembre 2025].

WORLD HEALTH ORGANIZATION; UNICEF. *Global report on assistive technology* [en línea]. Ginebra: WHO/UNICEF, 2022. ISBN 978-92-4-004115-7. Disponible en: <https://www.who.int/publications/i/item/9789240041157> [Consulta: 30 septiembre 2025].

WORLD WIDE WEB CONSORTIUM (W3C). *Web Content Accessibility Guidelines (WCAG) 2.1* [en línea]. 2018. Disponible en: <https://www.w3.org/TR/WCAG21/> [Consulta: 30 septiembre 2025].

AGRADECIMIENTOS

Este piloto se enmarca en las actividades de la **Cátedra de Brecha Digital y Discapacidad (DicaTIC)** de la **Universitat Politècnica de València (UPV)**, con el **apoyo financiero de la Generalitat Valenciana** y la colaboración del **Observatorio de Brecha Digital**. Asimismo, ha contado, con el **respaldo** de la **Fundación ONCE** y el **CERMI Comunidad Valenciana** para la validación de las pruebas y el desarrollo de acciones conjuntas en materia de accesibilidad. El equipo de **investigadores de la Cátedra DicaTIC**, ha constituido un apoyo fundamental en la **orientación metodológica** y en la consolidación de una visión interdisciplinar sobre inclusión y tecnología.

Este piloto refleja el compromiso de la UPV con los Objetivos de Desarrollo Sostenible (**ODS 9, 10 y 11**), promoviendo una **transformación digital inclusiva, ética y sostenible**, y expresa su agradecimiento a todas las personas y entidades colaboradoras que, con su **experiencia y sensibilidad hacia la accesibilidad universal**, han contribuido a enriquecer este trabajo.

Antonio Quirós Fons

Profesor titular de Derecho Internacional
Universidad Europea de Valencia
Cónsul Honorario de Croacia

[ACCEDE A LA GRABACIÓN](#)

La libertad de religión o de creencias es un derecho humano fundamental, consagrado en instrumentos internacionales como la Declaración Universal de los Derechos Humanos y el Pacto Internacional de Derechos Civiles y Políticos. En la actualidad, el ejercicio de este derecho se encuentra profundamente condicionado por las tecnologías digitales. La creciente digitalización de la expresión religiosa, la vida comunitaria y el culto plantea tanto oportunidades como riesgos: mientras que los entornos virtuales facilitan el acceso a recursos religiosos y la visibilidad de comunidades marginadas, también están sujetos a algoritmos opacos, vigilancia y censura.

Esta ponencia presenta algunos de los resultados ya publicados por el autor como resultado del estudio acerca del estado actual de la investigación en la intersección entre la libertad religiosa y las tecnologías digitales, con especial énfasis en la inteligencia artificial (IA) y los sistemas de vigilancia (Quirós Fons 2025). Se analizan los impactos de estas tecnologías en la moderación de contenidos, el reconocimiento facial, la vigilancia predictiva y la recopilación de datos biométricos, destacando sus implicaciones éticas, jurídicas y sociales. A través de una revisión crítica de la literatura existente, se identifican temas clave, vacíos conceptuales y oportunidades interdisciplinarias. El objetivo es ofrecer una hoja de ruta para futuras investigaciones que contribuyan a la protección y promoción de la libertad religiosa en un entorno digital en constante evolución.

Libertad religiosa y tecnología: algunos conceptos clave

La libertad religiosa es una piedra angular del derecho internacional de los derechos humanos, consagrada en el artículo 18 tanto de la Declaración Universal de los Derechos Humanos como del Pacto Internacional de Derechos Civiles y Políticos. Protege no solo el derecho a tener creencias, sino también a manifestarlas en la práctica, el culto y la observancia. En la era digital, este derecho se cruza cada vez más con los derechos digitales—como la libertad de expresión,

la privacidad y el acceso a la información—que son esenciales para salvaguardar la libertad religiosa en entornos en línea y mediados por la tecnología (Greiman 2021; Ünver 2024).

Derechos digitales y expresión religiosa

Los marcos de derechos digitales enfatizan la importancia del acceso equitativo a la tecnología, la protección contra la vigilancia y la capacidad de expresar creencias libremente en espacios digitales (Smuha 2025a). Estos derechos son especialmente relevantes dado que la expresión religiosa ocurre cada vez más en redes sociales, espacios de culto virtuales y a través del activismo digital.

Ética de la Inteligencia Artificial

La ética de la inteligencia artificial (IA), un campo en rápida evolución, aborda las implicaciones morales y sociales de la IA (Quirós-Fons 2025). Sus principios clave incluyen la equidad, la responsabilidad, la transparencia y el respeto por la dignidad humana (Floridi y Cowls 2019). Estos principios son vitales al considerar cómo los sistemas de IA afectan a individuos y comunidades religiosas, especialmente en áreas como la toma de decisiones automatizada (Osasona et al. 2024), la vigilancia (Almeida, Shmarko y Lomas 2022; Ünver 2024) y la moderación de contenidos (Ashraf 2021).

Enfoque pluralista

Elmahjub (2023) ofrece una intervención crítica al proponer un marco ético pluralista para la IA basado en la ética islámica. Critica el dominio de paradigmas centrados en Occidente en la ética de la IA, como una base moral secular y universalista, que marginaliza o excluye las tradiciones éticas religiosas y no occidentales. Otros paradigmas incluyen el individualismo liberal o la influencia tecnocrática y corporativa. Por ello, aboga por la inclusión de perspectivas religiosas y culturales diversas en el desarrollo y la gobernanza de tecnologías de IA. Este enfoque pluralista es esencial para garantizar que los sistemas de IA respeten los valores y derechos de las comunidades religiosas a nivel global (Elmahjub 2023).

Según este enfoque pluralista de la ética de la IA, el autor defiende la inclusión del razonamiento moral islámico junto a los marcos occidentales, introduciendo el concepto islámico de *maṣlaḥa* (bienestar o beneficio público) como principio ético central (Elmahjub 2023). Este concepto puede interpretarse de dos maneras complementarias. Primero, desde una perspectiva utilitarista o basada en el bienestar, apoya el desarrollo y la implementación de sistemas de IA que maximicen el beneficio social y contribuyan al bien común. Segundo, desde una perspectiva basada en el deber, enfatiza el cumplimiento de obligaciones morales y la voluntad divina, incluso en casos donde tales acciones no conduzcan al mayor beneficio. Esta interpretación dual permite un marco ético matizado que equilibra los resultados con los compromisos basados en principios.

Ejercicio digital de la libertad religiosa

McClure (2016) explora cómo las tecnologías digitales influyen en la creencia y el comportamiento religioso, introduciendo el concepto de “reparación religiosa” para describir cómo los individuos adaptan y personalizan su fe en entornos en línea. Por ejemplo, los usuarios pueden combinar textos religiosos tradicionales con reflexiones personales, arte digital o música contemporánea para crear publicaciones en Instagram o vídeos en TikTok que expresen su espiritualidad de una manera que resuene con su identidad y audiencia. Los hallazgos de McClure (2017) subrayan el impacto transformador de las plataformas digitales en la identidad y práctica religiosa.

Religión e Inteligencia Artificial

Singler y Watts (2024) han proporcionado una guía autorizada sobre la relación simbiótica entre religión e IA, demostrando cómo todas las religiones están interactuando actualmente con la inteligencia artificial. Madero, por su parte, profundiza en los conceptos teológicos básicos (2025). En el cristianismo, las aplicaciones impulsadas por IA ofrecen planes personalizados de estudio bíblico, recordatorios de oración e incluso pastores virtuales o chatbots que responden a consultas teológicas. Algunas iglesias han adoptado tecnología de reconocimiento facial para el seguimiento de asistencia y análisis de sentimientos durante los sermones. En el islam, las aplicaciones móviles utilizan IA para ayudar con los horarios de oración, la dirección de la Qibla y la identificación de alimentos halal, mientras que los académicos exploran sus implicaciones para la jurisprudencia islámica y la toma de decisiones éticas (Abu Bakar et alii 2024).

Otras tradiciones también están adoptando la IA de forma singular. En el budismo, monjes robóticos en Japón y China imparten enseñanzas, recitan sutras e interactúan con los visitantes del templo, fusionando la práctica espiritual con la innovación tecnológica (Trothen, Kwok y Lee 2024). El hinduismo utiliza herramientas de IA para facilitar rituales remotos, interpretar cartas astrológicas y permitir el darshan virtual mediante transmisiones en vivo y realidad aumentada (Khoirunisa et al. 2023). Mientras tanto, en el judaísmo, la IA apoya plataformas adaptativas de estudio de la Torá y contribuye a debates en curso sobre su papel en la toma de decisiones halájicas (Singler y Watts 2024).

Implicaciones en Derechos Humanos

Greiman (2021) destaca las implicaciones de la IA en los derechos humanos, señalando que el despliegue de sistemas de IA a menudo carece tanto de supervisión ética adecuada como de mecanismos de rendición de cuentas. Esto es particularmente preocupante para las minorías religiosas, que pueden verse afectadas de manera desproporcionada por sistemas de IA sesgados u opacos (Greiman 2021). La Unión Europea es la única región que ha desarrollado un marco legal integral para la inteligencia artificial, desde el influyente Reglamento General

de Protección de Datos (GDPR) hasta el recientemente adoptado Reglamento más conocido como IA Act (Smuha 2025b).

Áreas de investigación presentes en la actualidad

Observaciones Preliminares

Antes de presentar los hallazgos clave derivados del análisis temático, es importante destacar algunas observaciones preliminares sobre la metodología, el alcance geográfico y el contexto institucional de los estudios incluidos.

Metodología: La mayoría de los estudios emplearon análisis teóricos o conceptuales, combinando estos enfoques con análisis de políticas. Cabe destacar que no se identificaron estudios empíricos o cuantitativos en la muestra.

Alcance geográfico y contexto institucional: Un número reducido de estudios se centró en contextos nacionales específicos—como China y otros países individuales—o examinó grupos religiosos particulares, especialmente dentro de contextos islámicos. Sin embargo, la mayoría de los estudios abordaron cuestiones desde una perspectiva global o no especificaron un enfoque geográfico o institucional particular.

IA y moderación de contenidos

La moderación de contenidos impulsada por IA se ha convertido en un mecanismo central para gestionar el discurso en línea en las principales plataformas, concretamente en la gobernanza digital de la expresión religiosa (Ashraf 2021). Aunque estos sistemas están diseñados para detectar y eliminar contenido dañino, a menudo carecen del matiz cultural y religioso necesario para distinguir entre discurso de odio y expresión religiosa legítima (Bychkov y Prorvich 2021). Esto ha llevado a la eliminación excesiva o supresión de contenido religioso legítimo, particularmente de credos minoritarios (Popova 2024).

Las herramientas de moderación de IA producen con frecuencia falsos positivos (Popova 2024), etiquetando contenido religioso benigno como extremista u ofensivo. Esto es especialmente problemático en contextos multilingües y multiculturales donde la terminología religiosa puede ser malinterpretada por sistemas automatizados (Bontridder y Pouillet 2021). Además, estos sistemas, aunque destinados a combatir la desinformación, a menudo infringen la libertad de expresión y creencias debido a sus procesos de decisión opacos. La Ley de Servicios Digitales de la UE introduce nuevas obligaciones para que las plataformas respeten los derechos de los usuarios, lo que podría transformar el tratamiento del contenido religioso en línea (Quintais 2024; Quirós-Fons y García-Ull 2022).

El cambio hacia la moderación automatizada ha despolitizado la gobernanza del discurso en línea. Como muestra el análisis de Siapera (2021) sobre colonialidad y racismo en la moderación de IA, estos sistemas codifican normas culturales dominantes, marginando expresiones religiosas no occidentales.

Vigilancia y minorías religiosas

Las tecnologías de vigilancia, especialmente aquellas impulsadas por IA, están integradas en marcos legales que buscan mantener el orden público y la integridad territorial, pero pueden entrar en conflicto con las normas internacionales de derechos humanos, representando una amenaza directa para la libertad religiosa. Estas tecnologías han sido utilizadas por los estados para monitorear minorías religiosas bajo el pretexto de seguridad nacional.

Se documenta cómo herramientas impulsadas por IA, como el reconocimiento facial y de voz, y tecnologías generativas, apoyan la vigilancia estatal en China, restringiendo la actividad religiosa entre minorías étnicas y religiosas (Shumakova y Titova 2023). También se observan variaciones regionales en la intensidad de la vigilancia, con medidas más estrictas en áreas con poblaciones minoritarias significativas.

Además del ejemplo más citado—la vigilancia de musulmanes uigures en China (Shumatova and Titova 2023)—existen otros casos nacionales documentados donde se han utilizado herramientas de reconocimiento facial, seguimiento biométrico y policía predictiva para suprimir la práctica religiosa y la identidad (Greiman 2021; Ünver 2024). Como ya sucediera con el uso de deepfakes en conflictos bélicos (García-Ull and Quirós Fons 2022), la IA puede ser utilizada como arma para la desinformación y el control, exacerbando la marginación de grupos religiosos (Brenner et al. 2024). Greiman (2021) enfatiza la necesidad de marcos sólidos de rendición de cuentas para prevenir el uso indebido de la IA que infrinja los derechos humanos.

Este patrón no es aislado. En India, se han utilizado tecnologías de vigilancia para monitorear comunidades musulmanas, especialmente durante períodos de agitación política. En democracias occidentales, las poblaciones musulmanas también han sido desproporcionadamente objeto de programas de vigilancia antiterrorista, lo que genera preocupaciones sobre perfiles raciales y religiosos (Greiman 2021; Ünver 2024).

Incluso en sociedades democráticas fuera de regímenes totalitarios, los sistemas de IA pueden recopilar y procesar datos sensibles relacionados con las creencias y prácticas religiosas de los individuos, con implicaciones éticas sobre tecnologías y aplicaciones móviles utilizadas con fines religiosos (Ashraf 2021; Ty 2023). Estas prácticas no solo violan la privacidad, sino que también generan un efecto inhibitorio sobre la expresión religiosa, ya que los individuos pueden autocensurarse o evitar reuniones religiosas por temor a ser monitoreados. La normalización de la vigilancia en contextos religiosos pone en riesgo la libertad religiosa a nivel global.

Culto digital y práctica religiosa

La pandemia de COVID-19 aceleró la adopción del culto digital, con muchas comunidades religiosas recurriendo a transmisiones en vivo, servicios por Zoom y redes sociales para mantener prácticas comunitarias. Este cambio ha democratizado el acceso a los servicios religiosos, especialmente para quienes se encuentran en comunidades remotas o marginadas (Bingaman 2020).

Es importante distinguir este fenómeno del establecimiento del Camino del Futuro (WOTF), la primera iglesia oficialmente reconocida dedicada al culto de la inteligencia artificial. Fundada en 2015, WOTF representa un desarrollo significativo en el reconocimiento legal del transhumanismo como forma de religión, más que una adaptación digital de prácticas religiosas tradicionales (Vanoni 2020).

Sin embargo, el culto digital también introduce nuevos desafíos. Surgen preguntas sobre autenticidad, integridad ritual y presencia espiritual, particularmente en tradiciones que enfatizan la reunión física y la participación sacramental. Además, las plataformas digitales no son espacios neutrales; están gobernadas por algoritmos y políticas que pueden moldear el discurso religioso y su visibilidad (Ashraf 2021; Bingaman 2023; Campbell y Tsuria 2022). Las reuniones religiosas en línea suelen estar sujetas a monitoreo y recopilación de datos, lo que puede inhibir la expresión religiosa, especialmente en contextos autoritarios (Cobbe 2020). En este sentido, McClure habla en sus estudios de la remodelación que las plataformas digitales están operando en la práctica e identidad religiosa (2016, 2017).

También existe el problema de la exclusión digital. No todas las comunidades tienen acceso igualitario a internet o alfabetización digital, lo que puede agravar desigualdades existentes en la participación religiosa. Ty (2023) discute los desafíos éticos vinculados a estas tecnologías y enfatiza el papel potencial de las instituciones religiosas en el establecimiento de estándares éticos.

Sesgo algorítmico e identidad religiosa

El sesgo algorítmico ocurre cuando los sistemas de IA producen resultados socioculturales sistemáticamente injustos para ciertos grupos (Quirós-Fons 2023). En el contexto religioso, esto puede manifestarse de varias maneras: sistemas de reconocimiento facial que identifican erróneamente a personas con atuendos religiosos, o que funcionan mal en rostros no blancos y no masculinos; policía predictiva que apunta desproporcionadamente a minorías religiosas; o algoritmos de recomendación que amplifican estereotipos sobre ciertas religiones.

Estos sesgos a menudo provienen de datos de entrenamiento que subrepresentan o malinterpretan a los grupos religiosos, así como de la falta de diversidad entre los desarrolladores de IA. El resultado es un ecosistema digital que puede perpetuar la

discriminación y la marginación. Como solución, Elmahjub (2023) aboga por la inclusión de marcos éticos religiosos en el diseño de la inteligencia artificial para contrarrestar estos sesgos y promover tecnologías más inclusivas. Su trabajo destaca la importancia del pluralismo en la ética de la IA, especialmente en la protección de los derechos y la dignidad de las comunidades religiosas.

Conclusiones

A pesar del creciente interés, el campo sigue marcado por importantes vacíos: la escasa representación de perspectivas no occidentales, la falta de estudios empíricos, longitudinales y comparativos, una limitada implicación de las comunidades religiosas en el diseño tecnológico, y una colaboración interdisciplinaria insuficiente. A través de la revisión de la literatura existente y estudios de caso emergentes, también se han destacado direcciones prometedoras para futuras investigaciones, especialmente marcos interculturales.

Para garantizar que las tecnologías digitales respeten la libertad de religión o de creencias, se requiere un esfuerzo concertado por parte de académicos, responsables políticos y tecnólogos:

- Los académicos deben llevar a cabo investigaciones interdisciplinarias e inclusivas que se centren en las experiencias vividas por comunidades religiosas diversas. Esto incluye desarrollar marcos comparativos, realizar investigaciones participativas e integrar perspectivas teológicas en la ética digital.
- Los responsables políticos deben elaborar y aplicar normativas que protejan la libertad religiosa en los espacios digitales. Esto implica garantizar la transparencia y la rendición de cuentas en los sistemas de IA, prevenir abusos de vigilancia y promover el acceso equitativo a las plataformas digitales para todos los grupos religiosos.
- Los tecnólogos deben diseñar sistemas informados éticamente, culturalmente sensibles e inclusivos por defecto. Esto requiere colaboración con expertos en ética, juristas y comunidades religiosas para asegurar que la IA y las infraestructuras digitales no perpetúen sesgos ni exclusión.

En una era en la que las tecnologías digitales median cada vez más la vida pública, proteger la libertad religiosa no es solo un imperativo legal, sino también moral y social. El futuro de este derecho fundamental depende de nuestra capacidad colectiva para construir tecnologías—y sistemas de conocimiento—que sean justos, inclusivos y respetuosos de la dignidad humana.

Bibliografía

ABU BAKAR Azreen Zuhairi, NABGHAN Walid, ELHAEI Moslem, ABUL KHASSIM Zakiah, MUSTAFA Wan Azani, ABD AZIZ Noor Hazirah. "Comprehensive Review on Islamic Ethics and the Rise of Technology". En: *Journal of Advanced Research in Applied Sciences and Engineering Technology*. 2024. Vol. 51, nº 1, págs. 184–194. <https://doi.org/10.37934/araset.51.1.184194>

ALMEIDA, Denise, SHMARKO, Konstantin y LOMAS, Elizabeth. "The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks". En: *AI Ethics*. 2022. Vol. 2, págs. 377–387. <https://doi.org/10.1007/s43681-021-00077-w>

ASHRAF, Cameran. "Exploring the Impacts of Artificial Intelligence on Freedom of Religion or Belief Online". En: *The International Journal of Human Rights*. 2021. Vol. 26, n.º 5, págs. 757–791. <https://doi.org/10.1080/13642987.2021.1968376>

BINGAMAN, Kirk A. "Religious and Spiritual Experience in the Digital Age: Unprecedented Evolutionary Forces". En: *Pastoral Psychology*. 2020. Vol. 69, págs. 291–305. <https://doi.org/10.1007/s11089-020-00895-5>

BINGAMAN, Kirk A. "Religion in the Digital Age: An Irreversible Process". En: *Religions*. 2023. Vol. 14, n.º 1, artículo 108. <https://doi.org/10.3390/rel14010108>

BONTRIDDER, Noémi y POULLET, Yves. "The Role of Artificial Intelligence in Disinformation". En: *Data & Policy*. 2021. Vol. 3, artículo e32. <https://doi.org/10.1017/dap.2021.20>

BRENNER, David M. et al. "Technical and Religious Perspectives on AI Misinformation and Disinformation". En: *Social Science Research Network*. 2024. <https://doi.org/10.2139/ssrn.4360413>

BYCHKOV, Vasily V. y PRORVICH, Vladimir A. "The Role of Legal Algorithmic Language in the Identification, Disclosure, and Investigation of Extremist Crimes Committed Through Use of the Internet". En: *Russian Journal of Legal Studies*. 2021. Vol. 8, nº 3, págs. 27–40. <https://doi.org/10.17816/RJLS76068>

CAMPBELL Heidi, TSURIA Ruth. *Digital Religion: Understanding Religious Practice in Digital Media*. 2nd ed. Oxon: Routledge, 2022. ISBN 9780367272364. <https://www.routledge.com/Digital-Religion-Understanding-Religious-Practice-in-Digital-Media/Campbell-Tsuria/p/book/9780367272364>

COBBE Jennifer. "Algorithmic Censorship by Social Platforms: Power and Resistance". En: *Philosophy & Technology*. 2020.. <https://dx.doi.org/10.2139/ssrn.3437304>

ELMAHJUB Ezieddin. "Artificial Intelligence (AI) in Islamic Ethics: Towards Pluralist Ethical Benchmarking for AI". En: *Philosophy & Technology*. 2023. Vol. 36, nº 73. <https://doi.org/10.1007/s13347-023-00668-x>

FLORIDI Luciano, COWLS Josh. "A Unified Framework of Five Principles for AI in Society". En: *Harvard Data Science Review*. 2019. Vol. 1, nº 1. <https://doi.org/10.1162/99608f92.8cd550d1>

GARCÍA ULL, Francisco José y QUIRÓS-FONS, Antonio. "Inteligencia artificial y posverdad en tiempos de guerra". En: *Sociedad digital, comunicación y conocimiento: retos para la ciudadanía en un mundo global*. Madrid: Dykinson, 2022. Págs. 73–90. <https://dialnet.unirioja.es/servlet/articulo?codigo=8695366>

GREIMAN, V. "Human Rights and Artificial Intelligence: A Universal Challenge". En: *Journal of Information Warfare*. 2021. Vol. 20, nº 1, págs. 50–62. <https://www.istor.org/stable/27036518>

KHOIRUNISA, Ana, ROHMAN, Fathur, AZIZAH, Hilda Aulya, ARDIANTI, Devi, MAGHFIROH, Arinta Lailatul y NOOR, Aditya Muhammad. "Islam in the Midst of AI (Artificial Intelligence) Struggles: Between Opportunities and Threats". En: *SUHUF*. 2023. Vol. 35, nº 1, págs. 26–30. <https://doi.org/10.23917/suhuf.v35i1.22365>

MADERO, Kimberly A. "The Algorithm and the Almighty: Rethinking Omniscience, Suffering, and Salvation". En: *Religious Studies*. 2025. Págs. 1–3. <https://doi.org/10.1017/S0034412525000290>

McCLURE, Paul K. "Faith and Facebook in a Pluralistic Age: The Effects of Social Networking Sites on the Religious Beliefs of Emerging Adults". En: *Sociological Perspectives*. 2016. Vol. 59, nº 4, págs. 818–834. <https://doi.org/10.1177/0731121416647361>

McCLURE, Paul K. "Tinkering with Technology and Religion in the Digital Age: The Effects of Internet Use on Religious Belief, Behavior, and Belonging". En: *Journal for the Scientific Study of Religion*. 2017. Vol. 56, nº 3, págs. 481–497. <https://doi.org/10.1111/jssr.12365>

OSASONA, Femi, AMOO, Olukunle Oladipupo, ATADOGA, Akoh, ABRAHAMS, Temitayo Oluwaseun, FARAYOLA, Oluwatoyin Ajoke y AYINLA, Benjamin Samson. "Reviewing the Ethical Implications of AI in Decision Making Processes". En: *International Journal of Management & Entrepreneurship Research*. 2024. Vol. 6, nº 2, págs. 322–335. <https://doi.org/10.51594/ijmer.v6i2.773>

POPOVA, Svetlana S. "Use of Artificial Intelligence in the Activities of Religious Associations and Control Over Them". En: *Journal of Digital Technologies and Law*. 2024. Vol. 2, nº 1, págs. 101–122. <https://doi.org/10.21202/jdtl.2024.6>

QUINTAIS, João Pedro, ed. *From the DMCA to the DSA: A Transatlantic Dialogue on Online Platform Regulation and Copyright*. Berlín: Verfassungsbooks, 2024. <https://doi.org/10.17176/20240429-081042-0>

QUIRÓS-FONS, Antonio y GARCÍA ULL, Francisco José. "La inteligencia artificial como herramienta de la desinformación: deepfakes y regulación europea". En: *Los Derechos Humanos en la inteligencia artificial: su integración en los ODS de la Agenda 2030*. Cizur Menor: Thomson Reuters Aranzadi, 2022. Págs. 537–556. <https://dialnet.unirioja.es/servlet/articulo?codigo=8524831>

QUIRÓS-FONS, Antonio. "Estudio de estereotipos socioculturales en imágenes generadas por inteligencia artificial para el aprendizaje de los ODS y la mejora de las competencias digitales". En: *Sostenibilidad e internacionalización como pilares de vanguardia educativa*. Madrid: Dykinson, 2023. Págs. 278–298. <https://dialnet.unirioja.es/servlet/articulo?codigo=9234972>

QUIRÓS-FONS, Antonio, ed. *Propuestas educativas en la era de la IA. Regulación y uso ético*. Madrid: Dykinson, 2025. <https://doi.org/10.5281/zenodo.15805656>

QUIRÓS-FONS, Antonio. "Religious Freedom in the Age of AI and Surveillance: Mapping the Field and Charting Future Research Directions". En: *The Review of Faith & International Affairs*. 2025. <https://doi.org/10.1080/15570274.2025.2545140>

SHUMAKOVA, Natalya I. y TITOVA, Elena V. "Artificial Intelligence as an Auxiliary Tool for Limiting Religious Freedom in China". En: *Journal of Digital Technologies and Law*. 2023. Vol. 1, págs. 540–563. <https://doi.org/10.21202/jdtl.2023.23>

SIAPER, Eugenia. "AI Content Moderation, Racism and (de)Coloniality". En: *International Journal of Bullying Prevention*. 2021. Vol. 4, págs. 55–65. <https://doi.org/10.1007/s42380-021-00105-7>

SINGLER, Beth y WATTS, Fraser N., eds. *The Cambridge Companion to Religion and Artificial Intelligence*. Cambridge Companions to Religion. Cambridge: Cambridge University Press, 2024. <https://www.cambridge.org/core/product/867462871532038FDE79B253EA9A3C15>

SMUHA, Nathalie A., ed. *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Cambridge Law Handbooks. Cambridge: Cambridge University Press, 2025. <https://doi.org/10.1017/9781009367783>

SMUHA, Nathalie A. "AI, Law and Policy". En: *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*, editado por Nathalie A. Smuha. Cambridge Law Handbooks. Cambridge: Cambridge University Press, 2025. Págs. 131–258. <https://doi.org/10.1017/9781009367783.009>

TROTHEN, Tracy J., KWOK, Pui Lan y LEE, Boyung. "AI and East Asian Philosophical and Religious Traditions: Relationality and Fluidity". En: *Religions*. 2024. Vol. 15, n.º 5, artículo 593. <https://doi.org/10.3390/rel15050593>

TY, Rey. "Impact of AI-Powered Technology on Religious Practices and Ethics: The Road Ahead". En: *Religion and Social Communication*. 2023. Vol. 21, n.º 2, págs. 339–374. <https://doi.org/10.62461/RTY110823>

ÜNVER, H. Akin. *Artificial Intelligence (AI) and Human Rights: Using AI as a Weapon of Repression and Its Impact on Human Rights*. Think Tank. Parlamento Europeo, 2024. [https://www.europarl.europa.eu/thinktank/en/document/EXPO_IDA\(2024\)754450](https://www.europarl.europa.eu/thinktank/en/document/EXPO_IDA(2024)754450)

VANONI, Luca Pietro. "Deus ex machina. Intelligenza artificiale e libertà religiosa nel sistema costituzionale degli Stati Uniti". En: *Stato, Chiese e pluralismo confessionale*. 2020. N.º 15. <https://doi.org/10.13130/1971-8543/14048>

Noelia Liduina Rebón Rodríguez. Inteligencia artificial y acoso escolar.
Responsabilidad civil de los centros docentes en la era digital

Noelia Liduina Rebón Rodríguez

[ACCEDE A LA GRABACIÓN](#)

Resumen

La presente comunicación analiza el impacto de la inteligencia artificial (IA) en la configuración de la responsabilidad jurídica de los centros docentes ante los casos de acoso escolar y ciberacoso. El estudio parte del artículo 1903 del Código Civil —aplicable a los centros privados y concertados— y de los artículos 61 a 64 de la Ley Orgánica 5/2000, reguladora de la responsabilidad penal de los menores (LORPM), para aquellos supuestos en que los autores de los hechos tienen entre catorce y diecisiete años.

En el ámbito de los centros públicos, el marco jurídico se completa con el artículo 106.2 de la Constitución Española y los artículos 32 a 36 de la Ley 40/2015, de Régimen Jurídico del Sector Público, que regulan la responsabilidad patrimonial de la Administración. Desde esta perspectiva, se examina cómo las instituciones educativas deben responder civilmente por los daños ocasionados por los alumnos bajo su custodia, atendiendo al deber de vigilancia y supervisión que les corresponde.

El trabajo profundiza, además, en la relación entre tecnología y responsabilidad, analizando tanto el uso de la inteligencia artificial como herramienta de acoso —mediante la creación de *deepfakes*, la manipulación digital o la difusión no consentida de imágenes— como su utilización legítima por los centros docentes en la detección temprana y prevención de conductas de riesgo. Este doble enfoque permite valorar los retos que la digitalización y la automatización plantean al Derecho en materia de protección de menores y responsabilidad educativa.

Asimismo, se abordan las implicaciones del Anteproyecto de Ley Orgánica de protección de los menores en los entornos digitales (2025) y del Reglamento Europeo de Inteligencia Artificial (AI Act, 2024), junto con la aplicación práctica de diversas herramientas tecnológicas implementadas en centros escolares, como *Kanjo*, *el Termómetro del Acoso Escolar*, *Proyecto VIA*, *ReThink*, *b-resol*, *Moods* y *Kit ANAR*. Se incluyen también referencias a casos reales y resoluciones judiciales relevantes, entre ellas la sentencia del Juzgado de Menores de Badajoz

—caso de los “falsos desnudos” de Almendralejo (2024)— y distintos pronunciamientos de Audiencias Provinciales sobre la responsabilidad civil compartida entre progenitores y centros educativos.

Desde un enfoque garantista, se sostiene que la incorporación de sistemas algorítmicos en el ámbito educativo no sustituye el deber humano de vigilancia ni exime a los centros de su responsabilidad civil. Antes bien, impone una diligencia tecnológica reforzada, que debe ejercerse de manera compatible con la intimidad, la presunción de inocencia y el interés superior del menor. En consecuencia, la inteligencia artificial debe concebirse como un instrumento auxiliar al servicio de la tutela y la prevención del acoso, y no como un sustituto de la intervención humana.

Palabras clave

inteligencia artificial; acoso escolar; responsabilidad civil; centros docentes; menores; diligencia tecnológica; jurisdicción de menores; protección de datos; LORPM; LOPIVI; AI Act; Anteproyecto de Ley Orgánica de protección de los menores en los entornos digitales; ciberacoso; derecho educativo.

Abstract

This paper analyzes the impact of Artificial Intelligence (AI) on the configuration of the legal liability of educational institutions in cases of bullying and cyberbullying. The study draws upon Article 1903 of the Spanish Civil Code —applicable to private and publicly funded schools— and Articles 61 to 64 of Organic Law 5/2000, regulating the criminal liability of minors (LORPM), in those cases where the perpetrators are between fourteen and seventeen years of age.

In the context of public schools, the legal framework is complemented by Article 106.2 of the Spanish Constitution and Articles 32 to 36 of Law 40/2015, on the Legal Regime of the Public Sector, which govern the administrative liability of public authorities. From this perspective, the paper examines how educational institutions must respond in civil terms for the damages caused by students under their custody, in accordance with their duty of care and supervision.

The study also explores the relationship between technology and legal responsibility, analyzing both the use of artificial intelligence as a tool of harassment —through the creation of *deepfakes*, digital manipulation, or the unauthorized dissemination of images— and its legitimate application by schools in the early detection and prevention of risky behaviors. This dual perspective allows for an assessment of the challenges that digitalization and automation pose to the legal system in matters of child protection and educational liability.

Furthermore, the analysis addresses the implications of the Draft Organic Law on the Protection of Minors in Digital Environments (2025) and the European Artificial Intelligence Regulation (AI

Act, 2024), together with the practical implementation of various technological tools in schools, such as *Kanjo*, *El Termómetro del Acoso Escolar*, *Proyecto VIA*, *ReThink*, *b-resol*, *Moods*, and *Kit ANAR*. The paper also includes references to real cases and relevant judicial decisions, including the ruling of the Juvenile Court of Badajoz —the “false nudes” case of Almendralejo (2024)— and various judgments of Provincial Courts concerning shared civil liability between parents and educational institutions.

From a rights-based perspective, it is argued that the incorporation of algorithmic systems into the educational sphere neither replaces the human duty of supervision nor exempts schools from civil liability. On the contrary, it demands an enhanced form of technological diligence, to be exercised in a manner compatible with privacy, the presumption of innocence, and the best interests of the child. Consequently, artificial intelligence should be understood as an auxiliary instrument for protection and prevention against harassment, rather than as a substitute for human intervention.

Keywords

artificial intelligence; bullying; civil liability; educational institutions; minors; technological diligence; juvenile jurisdiction; data protection; LORPM; LOPIVI; AI Act; Draft Organic Law on the Protection of Minors in Digital Environments; cyberbullying; education law.

1. Introducción

La irrupción de la inteligencia artificial en el ámbito educativo ha transformado de manera profunda las dinámicas de aprendizaje, la gestión institucional y, en particular, los mecanismos de detección y prevención de la violencia entre iguales. Las herramientas basadas en algoritmos y análisis predictivos ofrecen hoy posibilidades inéditas para identificar conductas de riesgo, patrones emocionales o indicios de acoso escolar con una precisión antes inalcanzable. Sin embargo, esta misma capacidad tecnológica plantea interrogantes de gran calado jurídico y ético: ¿qué ocurre cuando los menores utilizan la inteligencia artificial para acosar o humillar a otros compañeros?, ¿en qué medida los centros educativos responden civilmente por los daños derivados de estos actos?, ¿qué responsabilidad asumen cuando aplican sistemas automatizados para prevenir el acoso y estos fallan, clasifican erróneamente o vulneran derechos fundamentales?

El objeto del presente estudio es analizar, desde una perspectiva jurídica y doctrinal, la responsabilidad civil de los centros docentes en los supuestos de acoso escolar mediado por inteligencia artificial, tanto cuando esta se utiliza como instrumento de agresión como cuando se aplica como herramienta de prevención. El análisis se estructura a partir de los fundamentos legales de la culpa *in vigilando*, consagrada en el artículo 1903 del Código Civil, y de la responsabilidad civil derivada de la jurisdicción de menores, recogida en los artículos 61 a 64 de la Ley Orgánica 5/2000, de 12 de enero, reguladora de la responsabilidad penal de los menores (LORPM).

El estudio tiene, además, una finalidad propositiva: reflexionar sobre la compatibilidad entre innovación tecnológica y tutela jurídica del menor, proponiendo un modelo de diligencia tecnológica reforzada que integre la inteligencia artificial en los centros docentes sin menoscabar la intimidad, la presunción de inocencia ni la dignidad del alumnado. Se pretende con ello ofrecer una visión equilibrada entre el progreso técnico y el respeto a los principios constitucionales, en particular el interés superior del menor (artículo 39 de la Constitución Española) y el derecho a la protección integral frente a la violencia (Ley Orgánica 8/2021, de Protección Integral a la Infancia y la Adolescencia frente a la Violencia – LOPIVI).

La metodología empleada combina un enfoque dogmático y jurisprudencial, con la exposición de casos reales, entre ellos el de Almendralejo (2024), en el que varios menores fueron condenados por la creación y difusión de falsos desnudos de compañeras mediante programas de inteligencia artificial, y la referencia a proyectos tecnológicos preventivos, como Kanjo, el Termómetro del Acoso Escolar, el Proyecto VIA, ReThink o b-resol, que evidencian las posibilidades y los límites del uso de algoritmos en el entorno escolar.

Este trabajo parte de una premisa fundamental: la inteligencia artificial puede constituir tanto una amenaza como una oportunidad. Su legitimidad depende del modo en que se integra en la

estructura jurídica y educativa. Cuando se emplea con supervisión humana, base normativa adecuada y respeto por los derechos fundamentales, la inteligencia artificial puede convertirse en un instrumento de protección y detección temprana del sufrimiento infantil. Pero cuando se usa sin control, o es manipulada por los propios alumnos para generar daño, puede dar lugar a una forma de violencia inédita, difusa y potencialmente devastadora.

1.2 Metodología y alcance del estudio

Este trabajo se apoya en: (i) análisis dogmático de la responsabilidad civil por hecho ajeno y por omisión; (ii) revisión sistemática de resoluciones judiciales relevantes en materia de acoso escolar y deber de vigilancia; (iii) examen de normativa europea y española aplicable a IA y protección de menores; y (iv) estudio de experiencias tecnológicas de prevención en centros. El criterio de selección de casos prioriza decisiones con impacto en culpa *in vigilando*, omisiones institucionales, y supuestos donde el daño se produce o amplifica en entornos digitales vinculados a la vida escolar. El enfoque es predominantemente preventivo-reparador, con propuestas de *lege ferenda*.

2. Transformación digital y nuevas formas de acoso escolar

El fenómeno del acoso escolar ha evolucionado en paralelo a la expansión tecnológica. La digitalización del entorno educativo ha trasladado los espacios de interacción social del aula al universo digital, multiplicando los escenarios de conflicto. Hoy el acoso no se limita al contacto físico o verbal, sino que se manifiesta en redes sociales, plataformas de mensajería, videojuegos y entornos virtuales donde la identidad y la imagen personal se convierten en vectores de agresión.

La inteligencia artificial generativa ha acentuado este fenómeno. Su capacidad para crear, modificar o difundir contenidos falsos con apariencia de veracidad ha dado lugar a nuevas modalidades de violencia. Entre ellas destacan los denominados *deepfakes* o “falsos desnudos”, imágenes o vídeos manipulados mediante redes neuronales (GANs) que simulan escenas de contenido sexual inexistentes. Este tipo de material, conocido como *undress AI*, se ha convertido en un instrumento de humillación pública y coacción emocional.

Un caso paradigmático fue el de Almendralejo (2024), donde quince menores fueron condenados por delitos de pornografía infantil y contra la integridad moral tras difundir imágenes falsas de compañeras generadas por inteligencia artificial. La sentencia del Juzgado de Menores de Badajoz (20 de junio de 2024) estableció un precedente al reconocer la gravedad autónoma del daño moral derivado de la manipulación digital, imponiendo medidas de libertad vigilada y formación en el uso responsable de las tecnologías de la información. El fallo puso de manifiesto que el acoso algorítmico, aunque mediado por tecnología, sigue siendo un acto humano que compromete la responsabilidad de quienes lo permiten o no lo previenen.

La proliferación de este tipo de conductas no se limita al ámbito español. Según la Fundación Internet Watch (IWF, 2025), en el primer semestre de ese año se detectaron 1.286 vídeos de abuso infantil generados por inteligencia artificial a nivel mundial, frente a solo dos en el mismo periodo del año anterior. En Estados Unidos, el Centro Nacional para Niños Desaparecidos y Explotados (NCMEC) registró 485.000 denuncias de contenido sexual infantil sintético en los seis primeros meses de 2025, frente a 67.000 en todo 2024 (*The New York Times*, 2025). La expansión global de este fenómeno evidencia la urgencia de articular respuestas legales que aborden el uso de inteligencia artificial como medio de agresión y no solo como herramienta de innovación.

En este contexto, los centros educativos no pueden permanecer ajenos. La culpa *in vigilando* adquiere una nueva dimensión cuando los actos lesivos se materializan a través de medios digitales o tecnológicos vinculados al entorno escolar. La responsabilidad del centro no se limita a lo que ocurre en el recinto físico, sino que se extiende a las actividades digitales derivadas de la convivencia educativa. Ignorar o minimizar los indicios de acoso en redes sociales o plataformas institucionales constituye, en términos jurídicos, una omisión relevante susceptible de generar responsabilidad civil.

Puente conceptual hacia la responsabilidad: La multiplicación de escenarios digitales no altera la naturaleza del deber: la misma lógica de custodia que opera en el patio y el aula se proyecta sobre los espacios digitales razonablemente conectados a la convivencia escolar (grupos de mensajería de clase, plataformas educativas, redes donde se replica el conflicto). Este desplazamiento exige ajustar los estándares de previsibilidad, trazabilidad y respuesta diligente.

3. Fundamento jurídico de la responsabilidad civil de los centros docentes

La responsabilidad civil de los centros docentes encuentra su fundamento en el artículo 1903 del Código Civil, que establece que “los padres son responsables de los daños causados por los hijos que se encuentren bajo su guarda, así como los tutores y los que estén al cuidado de menores o incapacitados, mientras permanezcan bajo su autoridad o vigilancia”. Asimismo, el precepto dispone que “los maestros o directores de artes y oficios responden por los daños causados por sus alumnos mientras estén bajo su control o enseñanza”.

Este artículo consagra el principio clásico de la culpa *in vigilando*, trasladado al ámbito educativo como un deber jurídico de prevención y control. La jurisprudencia ha interpretado de forma constante que los centros escolares actúan como guardadores de hecho durante el tiempo en que los menores se hallan bajo su cuidado (STS 22 de diciembre de 2012, AP Madrid). En virtud de ello, se les impone la obligación de evitar comportamientos dañinos tanto hacia terceros como entre los propios alumnos.

Esta responsabilidad es objetiva o cuasi objetiva, en tanto que no depende de la prueba de una culpa directa del centro, sino de la existencia de una relación de custodia y de la omisión de las medidas razonables para prevenir el daño. Solo la acreditación de una diligencia extraordinaria puede eximir de responsabilidad. Como señala Rivero Hernández (2019), la exoneración requiere demostrar una “vigilancia constante, proporcionada y adecuada al riesgo previsible”.

Con la incorporación de la inteligencia artificial en los centros, esta obligación se amplía. La adopción de sistemas algorítmicos de control o análisis emocional no libera a la institución de su deber de custodia, sino que introduce nuevas exigencias de diligencia tecnológica: verificar la idoneidad de los sistemas, garantizar su trazabilidad y respetar la confidencialidad y los derechos de los menores (Gascón Marcén, 2023).

El uso de herramientas algorítmicas para vigilar comportamientos o detectar acoso exige cumplir con los principios de proporcionalidad y transparencia recogidos en el Reglamento General de Protección de Datos (Reglamento UE 2016/679) y en la Ley Orgánica 3/2018 de Protección de Datos Personales y Garantía de los Derechos Digitales (LOPDGDD). La información generada por estos sistemas puede afectar a la reputación o el bienestar emocional de los alumnos, por lo que su implementación sin garantías podría suponer una vulneración de los derechos reconocidos en el artículo 18 de la Constitución Española.

La responsabilidad del centro se extiende, por tanto, tanto a las acciones como a las omisiones tecnológicas. La falta de adopción de medidas preventivas disponibles —como sistemas de alerta emocional o protocolos de detección digital— puede ser calificada como negligencia institucional, del mismo modo que un uso desproporcionado o intrusivo de tecnologías de control podría vulnerar derechos fundamentales. De ahí que el centro deba encontrar un equilibrio entre prevención y respeto a la privacidad, aplicando una ética de la vigilancia garantista (De Miguel Asensio, 2024).

El fenómeno del acoso escolar ha evolucionado en paralelo a la expansión tecnológica. [...] Ignorar o minimizar los indicios de acoso en redes sociales o plataformas institucionales constituye, en términos jurídicos, una omisión relevante susceptible de generar responsabilidad civil.

En suma, el artículo 1903 del Código Civil impone a los centros una obligación de resultado en materia de seguridad y convivencia: deben garantizar que los menores bajo su guarda no sufran ni causen daño durante el tiempo en que se hallen bajo su custodia, incluso en los entornos digitales asociados a la actividad escolar. La inteligencia artificial, bien utilizada, puede reforzar ese deber; mal empleada, puede convertirse en fuente de nueva responsabilidad.

3.1 Nexo causal y atribución en entornos digitales

En la práctica, el nexo causal se compone de dos planos:

Causalidad fáctica ampliada: que los hechos se inicien fuera del recinto o del horario no rompe per se la cadena causal cuando existe continuidad escolar (relaciones de clase, uso de canales del centro, conocimiento previo de indicios).

Imputación normativa: la atribución al centro se justifica cuando, pese a señales objetivas (avisos de familias, reportes de alumnado, detecciones de herramientas institucionales), no se activan protocolos razonables (prevención, comunicación, protección, documentación).

Este doble plano permite compatibilizar la autonomía de las conductas de los menores con la posición de garante del centro, evitando tanto la sobreimputación como la impunidad por omisión.

3.2 Deberes de diligencia tecnológica

La implantación de herramientas de inteligencia artificial en los centros educativos exige una diligencia tecnológica reforzada, es decir, un conjunto de precauciones que aseguren que el uso de estas tecnologías respeta los derechos de los menores y se ajusta a las obligaciones legales de vigilancia, protección y transparencia.

A continuación se explican los principales deberes que los centros deben cumplir para actuar con responsabilidad.

a) Idoneidad y proporcionalidad: elección de soluciones ajustadas a la finalidad protectora

Cuando un centro decide utilizar inteligencia artificial, debe hacerlo con un propósito legítimo y concreto, como la detección del acoso o del malestar emocional entre el alumnado. El principio de idoneidad implica que la herramienta elegida realmente sirva para esa finalidad, y que haya evidencia técnica o científica de su eficacia. Por su parte, la proporcionalidad exige que el medio utilizado no sea más invasivo de lo necesario.

Por ejemplo, si un objetivo puede alcanzarse mediante encuestas anónimas o aplicaciones emocionales sencillas, no sería proporcionado instalar sistemas de reconocimiento facial o analizar mensajes privados. La idea es usar la tecnología adecuada para proteger, no para vigilar.

Además, esta evaluación debe mantenerse en el tiempo: los sistemas deben revisarse periódicamente para confirmar que siguen siendo útiles, seguros y no excesivos.

Ejemplo:

Un colegio estudia varias aplicaciones para detectar acoso. Finalmente elige una que permite al alumnado expresar su estado de ánimo mediante iconos, descartando otra que utilizaba imágenes faciales. Con ello garantiza la finalidad protectora sin invadir la intimidad de los menores.

b) Evaluación de impacto en protección de datos (DPIA)

Antes de aplicar una herramienta de IA que trate datos personales de menores, el centro tiene la obligación de realizar una Evaluación de Impacto en Protección de Datos (DPIA). Este documento sirve para identificar los riesgos que la tecnología puede generar (por ejemplo, errores, sesgos o vulneraciones de privacidad) y establecer medidas para reducirlos.

El análisis debe describir cómo se obtienen y gestionan los datos, valorar si la herramienta es necesaria y proporcional, prever los riesgos posibles y definir medidas de seguridad como el cifrado, el control de accesos o la supervisión humana.

Ejemplo:

Antes de activar una aplicación que detecta lenguaje ofensivo, el centro realiza una DPIA y concluye que podrían producirse falsos positivos. Por eso, decide que todas las alertas deberán ser revisadas por el equipo de convivencia antes de tomar cualquier decisión.

c) Transparencia: información a familias y alumnado

La transparencia es una condición indispensable para que el uso de la inteligencia artificial sea legítimo. Las familias, el alumnado y el personal docente deben saber qué sistema se usa, para qué sirve, qué datos recopila y quién los gestiona.

El centro debe explicar de forma sencilla la finalidad del sistema, el tipo de información que analiza, las medidas de seguridad y los derechos que asisten a los interesados. Además, tiene que mantener registros claros de cada alerta o decisión generada por la IA, indicando quién la revisó y qué resultado tuvo. Esa trazabilidad demuestra que el proceso ha sido controlado y no automatizado de manera ciega.

Ejemplo:

Un instituto publica en su web una guía sobre su herramienta de detección temprana de acoso. Explica cómo funciona, quién puede ver las alertas y cómo las familias pueden solicitar información o ejercer sus derechos. Cada caso queda documentado con fecha, responsable y resolución.

d) Gobernanza de datos: minimización, seguridad y tiempo de conservación

La gobernanza de datos significa gestionar la información de manera responsable y segura. En los centros escolares esto es especialmente importante porque se trata de datos de menores.

Primero, se debe cumplir el principio de minimización, recogiendo solo los datos estrictamente necesarios. En segundo lugar, se deben aplicar medidas de seguridad, como el cifrado o los accesos restringidos, para evitar filtraciones. Y finalmente, se debe respetar el principio de

limitación temporal, eliminando los datos cuando ya no sean necesarios para la finalidad preventiva.

Ejemplo:

Un centro conserva los registros de alertas durante tres meses. Pasado ese tiempo, los elimina automáticamente, salvo en casos graves en los que la información se incorpore a un expediente formal.

e) Supervisión humana y posibilidad de impugnación

La tecnología no puede reemplazar el criterio humano. Toda alerta generada por un sistema de inteligencia artificial debe ser revisada por personas cualificadas antes de adoptar medidas sobre un alumno o alumna.

Asimismo, las familias deben tener la posibilidad de solicitar una revisión o impugnar una decisión, y el centro debe explicar de manera comprensible los motivos que llevaron a la alerta o a la intervención. De este modo se garantiza que la tecnología sirva como apoyo a la educación, y no como sustituto del juicio profesional o del acompañamiento humano.

Ejemplo:

Una aplicación detecta que un estudiante se aísla del grupo. Antes de activar el protocolo de acoso, el tutor investiga la situación y descubre que el alumno estaba enfermo. La alerta se corrige y se archiva, dejando constancia de la revisión.

f) Formación de equipos directivos y docentes

La formación del personal educativo es una parte esencial de la diligencia tecnológica. Los docentes y directivos deben conocer los riesgos de la inteligencia artificial, sus límites legales y las buenas prácticas para su uso responsable.

Es necesario formar en tres ámbitos principales:

Jurídico: derechos del menor, deber de vigilancia y normativa de protección de datos.

Técnico: interpretación de alertas, gestión de errores o falsos positivos.

Comunicativo y ético: trato adecuado con las familias y los alumnos, preservación de la privacidad y respeto a la dignidad personal.

La formación debe ser continua, práctica y evaluable, y actualizarse cada vez que cambien las herramientas o la normativa.

Ejemplo:

Cada año, el centro organiza talleres sobre IA y convivencia, y simulacros trimestrales para

practicar la gestión de alertas. Los resultados muestran mejoras en la rapidez de respuesta y en la calidad de las intervenciones.

g) Síntesis final

El cumplimiento de estos seis principios —idoneidad, proporcionalidad, evaluación de impacto, transparencia, gobernanza, supervisión y formación— representa el núcleo de la diligencia tecnológica reforzada.

Aplicarlos de manera real y documentada protege al alumnado y demuestra que el centro ha actuado con prudencia, responsabilidad y respeto a la legalidad. En consecuencia, su observancia puede atenuar o incluso exonerar la responsabilidad civil del centro conforme al artículo 1903 del Código Civil y al marco de la LOPIVI.

4. La inteligencia artificial como instrumento de acoso escolar

La evolución tecnológica ha modificado radicalmente la naturaleza del acoso escolar, extendiendo su alcance más allá de los muros del aula. El acoso ya no requiere presencia física ni contacto directo: puede manifestarse mediante insultos en redes sociales, llamadas intimidatorias, difusión de contenidos falsos o la manipulación de imágenes generadas por inteligencia artificial. La velocidad de propagación de estos contenidos y su apariencia de veracidad amplifican el daño, lo perpetúan en el tiempo y dificultan su reparación.

Las tecnologías de inteligencia artificial generativa, en particular las redes neuronales de tipo *Generative Adversarial Networks (GANs)*, permiten crear imágenes y vídeos sintéticos con un realismo que confunde incluso a los expertos. Estas herramientas, conocidas como *undress AI* o *deepfake generators*, posibilitan la alteración de fotografías inocuas de menores para simular desnudos o situaciones de contenido sexual, constituyendo una forma grave de violencia digital.

El caso de Almendralejo (Badajoz, 2024) se erige en paradigma de esta nueva realidad. Quince menores fueron condenados por delitos de pornografía infantil y contra la integridad moral al haber manipulado, mediante programas de inteligencia artificial, fotografías de compañeras de instituto para difundir falsos desnudos en redes sociales. La resolución judicial, dictada tras acuerdo entre las partes, impuso medidas de libertad vigilada, formación en igualdad y uso responsable de las tecnologías, subrayando que el objetivo del proceso no era la criminalización, sino la reeducación digital y afectiva de los infractores. El fallo reconoció que las imágenes, pese a ser sintéticas, producían un daño real y profundo a las víctimas, en forma de ansiedad, estigmatización y vulneración de su dignidad (Juzgado de Menores de Badajoz, 2024).

Este tipo de conductas encuentra ecos en la jurisprudencia internacional. En Estados Unidos, el Centro Nacional para Niños Desaparecidos y Explotados (NCMEC) reportó en 2025 más de 485.000 denuncias de imágenes generadas por IA con contenido sexual infantil, frente a 67.000 en todo 2024 (*The New York Times*, 2025). Estas cifras ponen de relieve la magnitud del problema y la insuficiencia de los marcos normativos tradicionales para abordar delitos donde el daño no proviene de un hecho físico, sino de la manipulación digital.

El Derecho español aún carece de una tipificación autónoma del acoso mediante inteligencia artificial, aunque la reforma penal de marzo de 2025 incorporó la sanción de uno a dos años de prisión para quienes utilicen IA para falsificar imágenes pornográficas sin consentimiento, especialmente cuando las víctimas sean menores. No obstante, desde la perspectiva civil, la responsabilidad de los centros educativos puede derivarse igualmente si estos hechos se producen en su entorno digital o físico, o si omiten medidas preventivas ante señales evidentes de riesgo.

La culpa por omisión resulta particularmente relevante. El centro educativo puede ser considerado responsable civilmente si, conociendo la existencia de un entorno digital donde se hostiga a un alumno —por ejemplo, un grupo de mensajería creado por estudiantes del centro—, no adopta acciones razonables de protección o comunicación a las autoridades competentes. La doctrina y la jurisprudencia han interpretado que la inacción ante indicios de acoso prolongado constituye una negligencia institucional. Así lo reconocieron las Sentencias de la Audiencia Provincial de A Coruña (2015) y de Alicante (2017), que declararon la responsabilidad civil de los centros por no haber activado protocolos de intervención frente a insultos reiterados en redes sociales y pintadas ofensivas que se prolongaron en el tiempo.

En consecuencia, la inteligencia artificial no solo ha multiplicado las formas de acoso, sino también los escenarios de responsabilidad. El daño digital —a diferencia del físico— deja una huella persistente y replicable, cuya difusión puede alcanzar dimensiones globales. La reparación civil en estos casos debe considerar no solo la indemnización económica, sino también la restitución moral y reputacional de la víctima, en consonancia con el principio de reparación integral recogido en la jurisprudencia del Tribunal Supremo (STS 17 de julio de 2019).

En supuestos de deepfakes o manipulación digital, la prueba del daño moral no exige acreditar veracidad del contenido, sino difusión y afectación a la dignidad; la reparación debe contemplar, además de la indemnización, medidas de desindexación, rectificación, apoyo psicológico y restauración reputacional.

5. La inteligencia artificial como herramienta de prevención del acoso

La misma tecnología que ha facilitado nuevas formas de violencia puede convertirse en un poderoso aliado para su detección y prevención, siempre que se utilice dentro de un marco ético y jurídico garantista. Diversas iniciativas desarrolladas en España y otros países demuestran que la inteligencia artificial, bien implementada, puede contribuir significativamente a la protección del alumnado y al cumplimiento del deber de vigilancia que pesa sobre los centros docentes.

5.1 Herramientas de intervención temprana

Entre los proyectos más destacados figura *Kanjo*, una aplicación de analítica emocional creada para detectar precozmente el sufrimiento infantil. Permite a los niños y niñas seleccionar dos veces al día cómo se sienten, generando datos anonimizados que son procesados mediante inteligencia artificial supervisada por profesionales de la psicología. Cuando el sistema detecta un patrón asociado al malestar, emite una alerta para que los equipos del centro actúen de forma preventiva. Con siete años de experiencia y más de 55.000 menores participantes, Kanjo ha logrado reducir en un 70% el tiempo medio de detección de malestar emocional y alcanzar una tasa de veracidad del 80% según el informe de impacto validado por UNICEF (2024).

El Termómetro del Acoso Escolar, implementado en Galicia, emplea algoritmos de predicción emocional para analizar el clima convivencial de los centros. Su objetivo es identificar desequilibrios que puedan anticipar conflictos, permitiendo a los equipos docentes intervenir antes de que el acoso se manifieste de forma abierta. Esta herramienta, basada en principios de anonimato y transparencia, evita vulnerar la confidencialidad de los menores, integrándose en los protocolos educativos de convivencia y bienestar.

El Proyecto VIA (Visión Artificial en el Aula), desarrollado en el Instituto Ribera del Tajo (Talavera de la Reina), constituye otro ejemplo relevante. Utiliza tecnologías de visión artificial, como *PyTorch*, *YOLO* y *MediaPipe*, para analizar la atención y la interacción de los alumnos durante las clases. El proyecto, enmarcado en la red nacional de centros de excelencia de Formación Profesional, permite detectar distracciones, desmotivación o signos de ansiedad, ofreciendo al profesorado información objetiva para mejorar la gestión emocional del aula. Si bien plantea interrogantes en materia de consentimiento y tratamiento de datos biométricos, su implementación controlada y con fines pedagógicos evidencia la capacidad de la IA para mejorar la convivencia y prevenir la exclusión.

5. 2 Otras iniciativas internacionales de prevención

Existen asimismo programas y herramientas complementarias que refuerzan la labor preventiva de los centros:

ReThink, desarrollada en Estados Unidos, detecta mensajes de ciberacoso antes de que sean enviados y propone al usuario reconsiderar su publicación. Ha demostrado reducir en un 93% la difusión de mensajes ofensivos entre adolescentes.

b-resol, canal seguro de comunicación anónima, permite a los alumnos denunciar situaciones de acoso, ciberacoso o violencia de género, conectando automáticamente con los responsables de convivencia del centro.

Moods, proyecto de intervención comunitaria, aborda los riesgos derivados del uso de las tecnologías en la adolescencia, integrando el componente emocional y familiar.

Kit contra el acoso escolar de la Fundación ANAR, que incluye materiales pedagógicos, canciones y recursos audiovisuales para trabajar la empatía y la prevención del acoso en el ámbito escolar y familiar.

Todas estas herramientas, al promover la participación activa de los menores y su capacidad de autodefensa emocional, refuerzan la responsabilidad institucional del centro y se alinean con la Ley Orgánica 8/2021 (LOPVI), que exige mecanismos efectivos para la detección y notificación del maltrato o acoso.

En términos jurídicos, el uso de estas tecnologías no exime al centro de su deber de vigilancia, pero sí puede servir como elemento exculpatario parcial, demostrando que se actuó con la diligencia exigible en la prevención del daño. En otras palabras, la no utilización injustificada de sistemas disponibles podría llegar a considerarse una omisión relevante, mientras que su uso sin garantías podría implicar una vulneración de derechos fundamentales. La prudencia jurídica exige, por tanto, equilibrio, transparencia y control humano efectivo.

6. Responsabilidad civil en la jurisdicción de menores

La Ley Orgánica 5/2000, de 12 de enero, reguladora de la responsabilidad penal de los menores (LORPM), contempla un sistema propio de responsabilidad civil en sus artículos 61 a 64, que opera de forma autónoma respecto del Código Civil. El artículo 61.3 establece que los padres, tutores, acogedores y guardadores legales o de hecho responderán solidariamente con el menor por los daños y perjuicios causados, “*por este orden*”. Este sistema, de carácter objetivo, persigue garantizar la reparación efectiva de las víctimas ante la frecuente insolvencia de los infractores.

El legislador busca con ello asegurar la función reparadora del derecho de daños, ampliando el ámbito de responsabilidad a quienes, por razón de guarda o custodia, tienen la capacidad o el deber de evitar el hecho lesivo. Se configura así una responsabilidad de imputación objetiva, moderable solo cuando los responsables no hubieran favorecido la conducta del menor con dolo o negligencia grave (art. 61.3 LORPM).

La cuestión controvertida se centra en determinar si los centros educativos pueden considerarse “guardadores de hecho” a los efectos de la norma. La doctrina mayoritaria sostiene que sí, en tanto ejercen funciones de vigilancia y tutela durante la jornada escolar. Así lo reconoció la Audiencia Provincial de Madrid (Sentencia de 22 de diciembre de 2012), al señalar que “para apreciar la responsabilidad del centro docente por los delitos o faltas cometidos por los alumnos, estos deben encontrarse bajo el control o vigilancia del profesorado, en el ejercicio de actividades docentes o complementarias”.

La Fiscalía General del Estado, en su Instrucción 10/2005 sobre tratamiento del acoso escolar desde el sistema de justicia juvenil, se pronuncia en el mismo sentido, subrayando que “los centros docentes tienen una indubitada responsabilidad en garantizar espacios seguros para que los menores cursen sus estudios libres de agresiones y vejaciones”.

La interpretación jurisprudencial de la expresión “*por este orden*” ha dado lugar a dos posiciones. Una primera, restrictiva, entiende que la responsabilidad se distribuye de forma excluyente, de manera que, si existen padres o tutores, los centros quedarían exentos (AP Castellón, Sentencia de 2 de febrero de 2010). La segunda, concurrente, defiende que deben responder solidariamente todos los sujetos que compartan deberes de vigilancia. Este criterio, sostenido por las Audiencias Provinciales de Cantabria (23 de diciembre de 2003) y La Rioja (8 de enero de 2015), resulta más coherente con el principio de protección integral del menor, pues reconoce la coexistencia de responsabilidades entre familia y escuela.

Desde un punto de vista de política jurídica, esta segunda tesis se ajusta mejor a la realidad educativa actual, donde la vigilancia del menor es una función compartida. Negar la responsabilidad de los centros supondría desvirtuar el principio de corresponsabilidad social en la protección de la infancia y dejar sin reparación adecuada muchos supuestos de acoso escolar ocurridos dentro o en conexión con el ámbito educativo.

Así, cuando los hechos delictivos ocurren durante la jornada escolar o en actividades organizadas por el centro, este puede ser declarado responsable civil solidario junto con los padres o tutores, incluso aunque no haya mediado negligencia directa. El fundamento de esta imputación radica en el deber institucional de garantizar entornos seguros, que hoy incluye no solo la seguridad física, sino también la digital.

La inteligencia artificial introduce aquí una nueva variable: la responsabilidad por omisión tecnológica. En la medida en que existan herramientas razonablemente accesibles para prevenir el acoso o monitorizar riesgos emocionales —como, por ejemplo: *Kanjo* o el Termómetro del Acoso Escolar—, la inacción del centro podría valorarse como falta de diligencia. Del mismo modo, el uso inadecuado o intrusivo de estas herramientas podría generar responsabilidad por vulneración de derechos fundamentales. En ambos casos, el

equilibrio entre vigilancia y libertad se convierte en el eje de la responsabilidad civil contemporánea en la jurisdicción de menores.

7. Responsabilidad patrimonial de la Administración educativa

En centros públicos, los hechos que deriven de funcionamiento normal o anormal del servicio pueden dar lugar, además, a responsabilidad patrimonial de la Administración (art. 106.2 CE; marco básico de la Ley 40/2015). Esta vía canaliza la reclamación cuando el daño trae causa en la organización y prestación del servicio educativo (p. ej., falta de protocolos, defectos en su activación, carencias notorias de recursos o de formación). Conviene, por ello, documentar: (i) protocolos vigentes; (ii) acciones desplegadas; (iii) cronología de alertas y respuestas; y (iv) coordinación con servicios sociales y fuerzas de seguridad.

8. Diligencia tecnológica y equilibrio jurídico

La consolidación de la inteligencia artificial como herramienta de apoyo en el ámbito educativo plantea un nuevo modelo de responsabilidad civil caracterizado por la exigencia de una diligencia tecnológica reforzada. Este principio, recogido de forma implícita en la doctrina de la culpa *in vigilando*, se ha visto progresivamente reforzado por el desarrollo normativo europeo y nacional en materia de protección de datos, seguridad digital y ética algorítmica.

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 inteligencia artificial (AI Act), constituye el eje normativo de esta nueva etapa. El texto europeo clasifica los sistemas de IA según su nivel de riesgo y exige que aquellos aplicados en ámbitos sensibles —como la educación, la gestión de menores o la evaluación psicológica— cumplan con estrictas garantías de transparencia, trazabilidad y supervisión humana. En el caso de los centros docentes, el AI Act impone obligaciones específicas de evaluación previa de impacto, documentación técnica y control del sesgo algorítmico, lo que obliga a replantear la noción clásica de diligencia del centro educativo.

En paralelo, el Anteproyecto de Ley Orgánica de protección de los menores en los entornos digitales (2025), impulsado en España, refuerza el deber de vigilancia institucional. Este texto normativo, en fase avanzada de tramitación, introduce un modelo de corresponsabilidad digital entre centros, familias y administraciones, imponiendo la obligación de garantizar entornos digitales seguros y protocolos de actuación ante la detección de riesgos tecnológicos. La norma establece además la necesidad de que toda herramienta tecnológica aplicada a menores sea evaluada en términos de proporcionalidad, finalidad legítima y respeto a los derechos fundamentales.

En consecuencia, el centro docente no puede limitarse a aplicar programas o aplicaciones de inteligencia artificial de manera acrítica. Su obligación de diligencia incluye la evaluación previa

de impacto sobre la privacidad (Data Protection Impact Assessment, DPIA), la verificación de las condiciones de consentimiento, la transparencia informativa hacia alumnos y familias, y la supervisión humana constante sobre las decisiones o inferencias que genere el sistema. La ausencia de estas garantías puede derivar en responsabilidad civil e incluso administrativa por infracción de la normativa de protección de datos (Reglamento UE 2016/679 y Ley Orgánica 3/2018).

La diligencia tecnológica reforzada no implica un deber ilimitado, pero sí exige una actitud activa, previsor y prudente frente a los riesgos asociados al uso de inteligencia artificial. El centro educativo actúa, en este contexto, como garante de la seguridad digital de los menores que se encuentran bajo su custodia. Este principio encuentra su fundamento en el interés superior del menor, consagrado en los artículos 2 y 11 de la Ley Orgánica 8/2021 (LOPVI), que obliga a las instituciones educativas a adoptar todas las medidas necesarias para proteger la integridad física, psíquica y moral de los menores frente a cualquier forma de violencia, incluida la derivada del entorno digital.

A su vez, el equilibrio jurídico entre innovación y protección se articula sobre tres ejes fundamentales:

1. Proporcionalidad. La implantación de tecnologías de inteligencia artificial debe ajustarse a la finalidad legítima de proteger al menor, evitando usos desmedidos o invasivos que puedan vulnerar su intimidad o libertad.
2. Transparencia y rendición de cuentas. Los centros deben documentar sus actuaciones tecnológicas y estar en condiciones de explicar los criterios utilizados por los sistemas algorítmicos.
3. Supervisión humana efectiva. Ningún sistema automatizado puede sustituir la capacidad de juicio del educador ni el acompañamiento emocional que requiere la función docente.

El equilibrio entre la diligencia exigible y la innovación necesaria exige comprender que el derecho no debe frenar la tecnología, pero sí dirigirla conforme a principios de prudencia y humanidad. Como señala De Miguel Asensio (2024), “la inteligencia artificial en el ámbito educativo sólo será legítima si actúa como extensión del criterio pedagógico y no como su reemplazo”.

El uso responsable de la inteligencia artificial en la escuela requiere, por tanto, una doble mirada: la jurídica, para garantizar la conformidad con el ordenamiento; y la ética, para asegurar que la tecnología sirva al desarrollo integral del menor y no a su estigmatización o control excesivo.

9. Conclusión

La inteligencia artificial ha transformado la estructura del riesgo en el ámbito educativo y, con ello, la extensión de la responsabilidad civil de los centros docentes. Los artículos 1903 del Código Civil y 61 a 64 de la LORPM mantienen su plena vigencia en la era digital, pero deben interpretarse a la luz de las nuevas realidades tecnológicas. El deber de custodia ya no se limita al espacio físico, sino que se proyecta hacia el ecosistema digital donde los menores conviven, aprenden y, en ocasiones, sufren agresiones invisibles.

El acoso escolar mediado por inteligencia artificial —sea mediante la creación de contenidos falsos, la manipulación de imágenes, el hostigamiento automatizado o el uso de algoritmos para amplificar la humillación— representa uno de los desafíos jurídicos más complejos de nuestro tiempo. Frente a ello, el Derecho debe ofrecer respuestas equilibradas que combinen la protección de las víctimas con la educación y reeducación de los infractores, especialmente cuando se trata de menores.

La responsabilidad civil de los centros docentes se extiende a estos supuestos no por una cuestión de culpa directa, sino por su posición de garantes. En su condición de guardadores de hecho, los centros están obligados a prevenir el daño, adoptar medidas de detección temprana y actuar ante los primeros indicios de acoso. Su inacción constituye una omisión relevante a efectos de responsabilidad, del mismo modo que un uso imprudente o invasivo de tecnologías de inteligencia artificial puede vulnerar derechos fundamentales.

La diligencia tecnológica reforzada, como nuevo estándar jurídico, exige que las instituciones educativas integren la inteligencia artificial de forma responsable, con evaluaciones de impacto, control humano y respeto a la privacidad. La innovación no exime del deber de custodia; al contrario, lo amplía y cualifica.

La corresponsabilidad digital —concepto emergente en el Anteproyecto de Ley Orgánica 2025— deberá consolidarse como pilar del derecho educativo contemporáneo. En ella confluyen la familia, la escuela y la sociedad, todas llamadas a garantizar que la tecnología sirva al bienestar y desarrollo del menor.

En definitiva, la inteligencia artificial puede y debe ser una aliada en la lucha contra el acoso escolar, siempre que se someta a los principios de legalidad, proporcionalidad, ética y humanidad. El reto jurídico de los próximos años será integrar estos principios en un marco normativo estable que permita a los centros docentes utilizar la tecnología como medio auxiliar de la vigilancia humana, nunca como sustituto de la misma.

Solo así la educación podrá mantener su función esencial: formar ciudadanos libres, conscientes y protegidos frente a las nuevas formas de violencia digital.

Bibliografía

Doctrina y estudios

Casadesus Ripoll, P. Inteligencia artificial y responsabilidad civil: perspectivas jurídicas y retos legislativos. *Revista de la Facultad de Derecho de México*, 2020, vol. 70, n.º 278-1, pp. 353–374. DOI: [10.22201/fder.24488933e.2020.278-1.77358](https://doi.org/10.22201/fder.24488933e.2020.278-1.77358).

Fundación Mutua Madrileña & Fundación ANAR. (2025). La opinión de los/as estudiantes: VII Informe de prevención del acoso escolar en centros educativos (Curso 2024-2025). Fundación Mutua Madrileña y Fundación ANAR. <https://www.anar.org/>

Biscaia Fernández, J. M.; Mohedano del Pozo, R. B.; Biscaia Fernández, C. J. La inteligencia artificial en la prevención de conductas suicidas: aspectos técnicos y consideraciones ético-legales. *Revista de bioética y derecho*, 2023, nº 59, pp. 181-203. ISSN 1886-5887. doi: 10.1344/rbd2023.59.42759. Disponible en: <https://revistes.ub.edu/index.php/RBD/article/view/42759>.

Rebón Rodríguez, N. L. (2024). Entre el aula y el algoritmo: el acoso escolar en la era de la Inteligencia Artificial. *Law & Trends*. Disponible en: <https://www.lawandtrends.com/noticias/penal/entre-el-aula-y-el-algoritmo-el-acoso-escolar-en-la-era-de-la-inteligencia-artificial-1.html> (Consulta: 8 octubre 2025).

Rebón Rodríguez, N. L. (2024). El acoso escolar en la Memoria de la Fiscalía General del Estado: primera inclusión estadística. *Law & Trends*. Disponible en: <https://www.lawandtrends.com/noticias/penal/el-acoso-escolar-en-la-memoria-de-la-fiscalia-general-del-estado-primer-inclusion-estadistica-1.html> (Consulta: 8 octubre 2025).

Normativa

España. Código Civil. Real Decreto de 24 de julio de 1889. *Gaceta de Madrid*, 25 de julio de 1889.

España. Ley Orgánica 5/2000, de 12 de enero, reguladora de la responsabilidad penal de los menores. *Boletín Oficial del Estado*, n.º 11, 13 de enero de 2000.

España. Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. *Boletín Oficial del Estado*, n.º 294, 6 de diciembre de 2018.

España. Ley Orgánica 8/2021, de 4 de junio, de Protección Integral a la Infancia y la Adolescencia frente a la Violencia (LOPVI). *Boletín Oficial del Estado*, n.º 134, 5 de junio de 2021.

Unión Europea. Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio, por el que se establecen normas armonizadas en materia de inteligencia artificial (AI Act). *Diario Oficial de la Unión Europea*, L 1689, 13 de junio de 2024.

Documentos preparatorios (sin valor normativo)

España. Anteproyecto de Ley Orgánica de protección de los menores en los entornos digitales. Ministerio de Juventud e Infancia, Madrid, 2025.

Jurisprudencia

Audiencia Provincial de A Coruña. Sentencia núm. 742/2015, de 15 de octubre de 2015. ECLI: ES:APC:2015:742. Disponible en: <https://www.poderjudicial.es/search/> (Consulta: 8 octubre 2025).

Audiencia Provincial de Alicante. Sentencia núm. 541/2017, de 10 de mayo de 2017. ECLI: ES:APA:2017:541. Disponible en: <https://www.poderjudicial.es/search/> (Consulta: 8 octubre 2025).

Audiencia Provincial de Cantabria. Sentencia núm. 1453/2003, de 23 de diciembre de 2003. ECLI: ES:APC:2003:1453.

Audiencia Provincial de Castellón. Sentencia núm. 231/2010, de 2 de febrero de 2010. ECLI: ES:APCS:2010:231.

Audiencia Provincial de La Rioja. Sentencia núm. 56/2015, de 8 de enero de 2015. ECLI: ES:APLR:2015:56.

Audiencia Provincial de Madrid. Sentencia núm. 9871/2012, de 22 de diciembre de 2012. ECLI: ES:APM:2012:9871.

Juzgado de Menores de Badajoz. Sentencia de 20 de junio de 2024 (caso “falsos desnudos” de Almendralejo). No publicada; referencia: Fiscalía de la Comunidad Autónoma de Extremadura.

Tribunal Supremo. Sentencia núm. 2429/2019, de 17 de julio de 2019. ECLI: ES:TS:2019:2429. Disponible en: <https://www.poderjudicial.es/search/> (Consulta: 8 octubre 2025).

Kaj Wik Siebert. From predicting the past to designing the future: A systems level simulation approach for children's services

Kaj Wik Siebert

Independent Consultant, Data & AI for Social Impact, kaj@k-si.com

[ACCEDE A LA GRABACIÓN](#)

Abstract

The United Kingdom's children's services system is in a state of perpetual crisis, characterized by unsustainable financial pressures, escalating demand, and profoundly poor outcomes for the vulnerable children and young people it is mandated to protect. In response, a prevailing technocratic consensus has championed the use of data-driven solutions, particularly predictive analytics to improve efficiency and target interventions. This paper critically deconstructs this paradigm, arguing that it is not only ethically fraught but also built upon a fundamental and unsolvable statistical fallacy: the Age-Period-Cohort (APC) identification problem. Through an analysis of historical technological failures, such as the UK's ContactPoint database and the US's Allegheny Family Screening Tool, this report demonstrates that predictive models inevitably amplify existing societal biases and create a dangerous veneer of objectivity while lacking statistical validity for long-term forecasting. As an alternative, this paper proposes a shift away from predicting individual risk and towards systems-level simulation. It details the principles of Markov-chain modelling, Agent-Based Modelling (ABM), and Social Digital Twins (SDTs) as tools for understanding systemic dynamics and testing policy interventions in a virtual "sandbox." A blueprint for a multi-scale children's services simulator is outlined, alongside suggestions for a robust ethical framework, grounded in UNESCO's recommendations and tailored to address the unique challenges of feedback loops in dynamic digital twin systems. The central thesis is that the goal should not be to predict the past with greater accuracy, but to design better, fairer, and more effective systems for the future.

Keywords: children's services; predictive analytics; simulation; artificial intelligence; ethics; digital twins; age-period-cohort problem; public policy

Section 1: The System Under Strain: A Crisis in UK Children's Services

One could argue that the children's social care system in the United Kingdom is failing. This is not a failure of intent or of the dedicated professionals on the front line, but a systemic failure, evident in a set of mutually reinforcing crises in finance, capacity, and most critically human outcomes. The system is buckling under the immense pressure of rising demand and complexity, leading to a reactive posture that prioritizes costly, late-stage interventions at the expense of preventative support.

1.1 The Economics of Failure: Spiralling Costs and Misaligned Spending

The financial state of children's services is unsustainable. Local authorities have been trapped in a cycle of overspending, exceeding their children's social care budgets every year since 2014. In the 2021/22 financial year, total spending reached £11.1 billion, representing a staggering 41.6% increase in real terms since 2009/10. This persistent overspend, which saw 46% of councils exceed their budgets by at least 20% in 2021/22, is not a sign of profligacy but of a system overwhelmed by demand it cannot control (CIPFA 2025).

This expenditure is not evenly distributed across the spectrum of need; it is heavily skewed towards acute, high-cost interventions. The primary driver of these rising costs is residential care for children with complex needs. Spending in this area has surged by 66.2% in real terms between 2015/16 and 2021/22 alone. This has created a critical and damaging strategic trade-off. To meet their statutory duties for children in care, local authorities have been forced to divert funds away from early intervention and preventative services. The consequences are stark: between 2009/10 and 2021/22, while spending on children in care rose by 49.4%, spending on vital preventative services like children's centres and support for children under five plummeted by 73.4%.

This financial dynamic creates a systemic "death spiral." The disinvestment in preventative services, which are designed to support families and mitigate the stresses that can lead to children entering care, directly fuels the rise in complex, high-cost acute cases. As more families reach a crisis point, demand for expensive residential and foster care placements increases. The soaring cost of these acute services then consumes an ever-larger portion of the budget, forcing councils to make even deeper cuts to the remaining preventative services to meet their legal obligations. This completes a vicious, self-reinforcing feedback loop of escalating costs and worsening outcomes. The system is structurally incapable of investing in the long-term solutions that would reduce demand because it is perpetually overwhelmed by the immediate cost of managing a demand it inadvertently helps to create through its own funding logic. The market for placements is now so strained that the Department for Education has categorized the risk of market failure as "critical to very likely," highlighting a fundamental structural fragility that threatens the system's core function.

1.2 The Human Cost: Rising Demand and Poor Life Outcomes

The financial crisis is a symptom of a deeper human crisis. Demand for services is increasing in both volume and complexity. In 2021/22, there were 650,270 referrals to children's social care, and the number of children in care reached 82,170, a 27% increase since 2009/10. A significant new driver of this demand is the rise in unaccompanied asylum-seeking children, who now account for 17.1% of all children starting a care episode and often present with complex trauma needs requiring intensive support (CIPFA 2025).

For the children and young people who experience this system, the long-term outcomes are devastatingly poor, revealing a state that often fails in its role as a corporate parent. The data paints a bleak picture across every key domain of adult life.

Health and Wellbeing: The health outcomes for care-experienced individuals are catastrophic. In England, they face a 62% higher risk of dying before the age of 75 compared to their peers who have not been in care. In Northern Ireland, the statistics are even more alarming: young adults with care experience are nine times more likely to die by suicide (UKRI 2025).

Education: The educational attainment gap is a chasm. In England, 65% of young people not in care achieve a grade 4 or above in English and Maths at Key Stage 4. For children who have been in the care system for more than 12 months, this figure is just 20% (UKRI 2025). This disadvantage persists into higher education, where care-experienced young people are around four times less likely to enter by the age of 22 and two and a half times more likely to drop out if they do manage to enrol (TASO 2025).

Employment and Economic Stability: The transition to adulthood is a 'cliff-edge' for many care leavers (Barnardo's 2021). Eight years after leaving school, 25% of care-experienced individuals in England had not been in employment or education, compared to just 4% of their peers. A staggering 77% had received benefits at some point in those eight years, versus 21% among all school leavers (UKRI 2025). National figures show that 40% of all care leavers aged 19-21 are not in education, employment, or training (NEET) (Barnardo's 2021).

Housing and Justice: A stable home is a foundation for a successful life, yet it is denied to many care leavers. The expectation of independence at 18 leaves many vulnerable, with one in three (35%) living independently by age 19-21, increasing their risk of homelessness (Barnardo's 2021). Furthermore, contact with the justice system is disproportionately high; children with care experience are, on average, over four times more likely to have a criminal conviction or police caution than their peers (UKRI 2025).

1.3 Innovating Under Pressure: Data Pioneers

In response to this systemic crisis, organizations such as Data to Insight, Coram and Social Finance have pioneering new models to improve service delivery, drive accountability, and achieve better value for money. Their work represents a sophisticated attempt to optimize the performance of the existing system through the strategic use of data, finance, and cross-sector collaboration.

While this work is laudable and necessary, it functions primarily as a high-level optimization layer on a fundamentally broken system architecture. The tools and financial models make the existing system more efficient, more accountable, and better at managing its resources. However, they do not, and cannot, change the underlying reactive nature of the system itself. An outcomes rate card can incentivize a provider to deliver a better home visiting service, but it does not alter the broader socioeconomic conditions that put a family in crisis. A demand forecasting tool can help a local authority budget more accurately for the number of foster placements it will need, but this is an act of managing the consequences of the system's structure, not changing the structure itself.

This reveals the profound limitations of performance management and financial innovation in the absence of a more fundamental redesign of the system's core logic. One can have the world's most sophisticated contract and the most accurate forecast, but if the only interventions available are late-stage and high-cost, one is merely optimizing the management of failure. This realization provides the crucial bridge to this paper's central argument: we do not just need to manage the current system better; we need to design a better system entirely.

Section 2: A Legacy of Failure: Technological Interventions in Child Welfare

The persistent crisis in children's services has created a powerful allure for technological solutions. The promise of data and algorithms to bring order to a complex and high-stakes domain is compelling. However, the history of large-scale technological projects in child welfare is a sobering chronicle of spectacular failures.

These projects have not only wasted vast sums of public money but have often exacerbated the very problems they were designed to solve, causing significant harm to vulnerable children and families. This section provides a critical analysis of two landmark cases: the UK's ContactPoint database and the US's Allegheny Family Screening Tool. These case studies are not an argument against technology per se, but rather serve as cautionary tales, illustrating the specific conceptual and ethical pitfalls that any new technological paradigm must rigorously avoid.

2.1 Case Study: The UK ContactPoint Database – A Failure of Centralized Surveillance

Conceived in the aftermath of the tragic death of Victoria Climbié, the ContactPoint database was a monumental IT project with a seemingly simple goal: to ensure that professionals working with children could identify and communicate with one another. The plan was to create a single, centralized database holding basic identifying information but crucially no case details, for all 11 million children in England. If a professional had a concern, they could look up the child and see which other services (e.g., school, GP, social worker) were also involved in their life.

After costing £224 million to set up, the project was summarily scrapped by the incoming Coalition Government in 2010 before it was ever fully operational. Its failure was the result of a confluence of deep-seated flaws. (*Wikipedia* 2025, 'ContactPoint')

Privacy and Civil Liberties: From its inception, ContactPoint was condemned by a broad coalition of civil liberties groups, medical bodies, and academics as a disproportionate and dangerous invasion of family privacy. The British Medical Association raised concerns about the erosion of doctor-patient confidentiality, while groups like Liberty warned that it amounted to a state "surveillance tool" that would normalize the monitoring of every child and family, regardless of risk.

Massive Security Risks: The proposal to grant access to an estimated 330,000 users created an unmanageable security risk. Public and political confidence was shattered by the 2007 scandal in which HM Revenue and Customs lost the unencrypted personal data of 25 million child benefit recipients, an event that forced a major delay and security review of the ContactPoint project. There were credible and persistent fears that the database could be infiltrated or abused by those seeking to identify and harm vulnerable children.

A Fundamental Conceptual Flaw: The project's most profound failure was its flawed premise. It assumed that the primary barrier to effective child protection was a lack of information sharing, solvable by a central directory. Critics argued forcefully that this was an oversimplification. The real challenges lay in professional judgment, resource allocation, and the complex, relational nature of social work. There was a significant risk that the system would drown serious cases in a sea of low-level, decontextualized "concerns," while simultaneously undermining the authority and responsibility of parents and skilled professionals to manage risk appropriately. ContactPoint was a classic example of a technological solution being applied to a deeply human and systemic problem.

2.2 Case Study: The US Allegheny Family Screening Tool (AFST) – A Failure of Predictive Analytics

If ContactPoint represented the failure of centralized data aggregation, the Allegheny Family Screening Tool (AFST) represents the failure of the next evolution in techno-solutionism: predictive risk modelling. Deployed in Allegheny County, Pennsylvania, the AFST is an algorithmic tool designed to analyse calls made to the child protection hotline. It generates a "risk score" from 1 to 20, intended to help human call screeners decide whether a case warrants a formal investigation. (Allegheny County, n.d.)

Despite claims by the county of its ethical appropriateness and superior accuracy compared to human judgment alone, the AFST has been subject to withering criticism (Eubanks 2018) for being statistically flawed, racially biased, and operationally ineffective.

Flawed Proxies and Statistical Invalidity: The model's primary failure lies in what it measures. It was not trained on verified instances of child abuse or neglect. Instead, it was trained to predict proxies for harm, such as the likelihood of a "re-referral" to the service within two years or a child's subsequent placement into foster care. This is a critical statistical flaw. Re-referrals can be driven by a host of factors unrelated to child harm, such as contentious custody disputes. By using these flawed proxies, the model is not predicting risk of harm; it is predicting future system contact. This creates a tool that is not measuring what it purports to measure.

Racial Bias and Amplification: The AFST operates within a system that is already marked by profound racial disproportionality. In the US, Black children are investigated at nearly double the rate of white children. The historical data used to train the AFST is a direct reflection of this systemic bias. The algorithm, in turn, learns and codifies these biases. The result is that Black children are approximately 1.7 times more likely than white children to be assigned a "high-risk" score by the AFST (Samant et al. 2023).

This creates a pernicious feedback loop: the tool's biased predictions lead to increased surveillance of Black communities, which generates more system data on those communities, which is then used to retrain the model, further confirming and amplifying the original bias.

The Veneer of Objectivity: The tool's output, *a single number*, creates a powerful illusion of scientific precision and objectivity. It suggests an impartial, evidence-based assessment of risk. However, the thresholds that separate "low," "medium," and "high" risk scores are not statistically derived. They are arbitrary policy decisions set by the agency and have been changed at least four times since 2018 (Samant et al. 2023). This practice masks subjective, high-stakes judgments behind a seemingly empirical and unassailable number.

Lack of Efficacy: One of the key justifications for the tool was that it would help screen out low-risk referrals, reducing the number of families subjected to unnecessary and often traumatic investigations. However, data indicates that overall screen-in rates in Allegheny County did not

decrease following the tool's implementation, suggesting it failed to deliver on this core promise.

The failures of ContactPoint and the AFST are not disconnected events. They represent the collapse of two successive, flawed techno-philosophies that have dominated attempts to reform child welfare. ContactPoint embodied the "data aggregation" fallacy: the belief that if we simply have all the data in one place, we can connect the dots and prevent harm. It failed because it stripped data of its context, overwhelming professional judgment with noise.

The AFST embodies the "predictive analytics" fallacy: the belief that if we apply a clever algorithm to historical data, we can predict future harm. It failed because the historical data was a product of systemic bias, which the algorithm inevitably learned and amplified, creating a more efficient and seemingly objective engine for discrimination.

The progression from aggregation to prediction appears sophisticated, but it doubles down on the same fundamental error: the belief that risk is a static, individual attribute that can be identified through discrete data points. It ignores the reality that risk is a dynamic, emergent property of a complex system of relationships between a family, its community, and its socioeconomic environment. This history demonstrates that the next evolution in the use of technology cannot be a "better" predictive model. It requires a complete paradigm shift away from individual profiling and towards a systems-level perspective.

Section 3: The Temporal Chasm: Why Prediction Fails in Complex Social Systems

The failures of predictive tools like the Allegheny Family Screening Tool are not merely isolated incidents of poor implementation. They are the inevitable symptoms of a deep, foundational flaw in the entire predictive analytics paradigm as applied to long-term human outcomes. This problem, known as the Age-Period-Cohort (APC) identification problem, creates a *temporal chasm* between the world in which the data was generated and the world in which the prediction is made, rendering such forecasts statistically invalid and ethically indefensible.

3.1 The Age-Period-Cohort (APC) Identification Problem: A Statistical Impasse

The Age-Period-Cohort (APC) identification problem is a fundamental challenge in the social sciences that has stymied longitudinal research for nearly a century (Bell and Jones 2015). It arises from the perfect linear dependence between the three temporal dimensions that shape all human experience: age, historical period, and birth cohort.

Age = Period – Cohort

For any individual, their age is perfectly determined by the current calendar year (period) and their year of birth (cohort). This perfect collinearity makes it statistically impossible to

disentangle the independent, linear causal effects of these three distinct forces on any given outcome.

For example, if we observe a rising rate of mental health issues among teenagers, the APC problem makes it impossible to determine, through standard regression techniques, whether this is due to:

- An **Age Effect**: A biological or developmental process inherent to being an adolescent.
- A **Period Effect**: A societal-level event or trend affecting everyone at a specific point in time, such as the rise of social media, economic recession, or a global pandemic.
- A **Cohort Effect**: A unique characteristic shared by a group of people born around the same time (e.g., “Millennials” or “Gen Z”), shaped by their unique formative experiences.

Because of the linear dependency, any one of these linear effects can be expressed as a combination of the other two. This means an infinite number of different combinations of age, period, and cohort effects can produce the exact same observable data.

As the statistician Norval Glenn noted, the continued search for a purely statistical technique to solve this problem is "one of the most bizarre instances in the history of science of repeated attempts to do the logically impossible" (Glenn 2005). The problem is not a limitation of our data or our models; it is a logical property of time itself. It cannot be solved with a larger dataset or a more powerful algorithm. Any model that claims to have "solved" the identification problem has invariably done so by imposing a constraint (e.g., assuming the effect of one period is equal to another) that is theoretically arbitrary and impossible to defend.

3.2 The Illusion of Objectivity: APC and the Invalidity of Long-Term Prediction

The APC problem is not an obscure academic puzzle; it is the theoretical linchpin that explains why predictive models in child welfare are destined to fail. These models, built on archives of historical administrative data of varying quality, operate on the implicit and deeply flawed assumption that the statistical correlations observed in the past will hold true for the future. The most critical indicators of the system's failure, such as poor long-term health, educational attainment, or involvement with the criminal justice system, are not observable until many years, often decades, after a child has left care. Consequently, any dataset of adult outcomes is, by definition, a record of a cohort of children who experienced the care system in a vastly different historical period.

The social, economic, and policy environment (the Period effects) they navigated is not the environment of today, and their own generational experiences (the Cohort effects) are distinct. Using the outcomes of a child from the 2000s to predict the trajectory of a child in the 2020s is

not just a matter of biased data; it is a fundamental comparison of two different social realities, making the prediction statistically indefensible.

This creates a "temporal chasm" between the context in which the model's "knowledge" was created and the context in which it is being applied. The model is not truly predicting the future; it is projecting a statistically confounded and contextually obsolete past onto the present.

The result is a tool that provides a dangerous "veneer of objectivity." It presents its risk score as a scientific, impartial fact, when it is an artifact of an unsolvable statistical problem, systematically biased data, and arbitrary policy choices. This elevates the critique of tools like the AFST from an empirical observation that "this specific tool is biased" to a theoretical proof that "any tool of this type, when used for long-term social prediction, is statistically invalid at its core."

The goal of creating a purely data-based, fair, and unbiased predictive tool for child welfare is not just technically difficult; it is logically incoherent. This realization provides the ultimate justification for abandoning the predictive paradigm entirely in favour of an approach that acknowledges and embraces complexity and uncertainty.

Section 4: From Prediction to System Simulation

The deconstruction of the predictive paradigm necessitates a constructive alternative. If we cannot attempt to predict the future of an individual child, we must instead shift our focus to understanding the system that shapes their future.

This requires a new class of tools designed not for prediction, but for exploration, learning, and policy design. This section introduces a suite of simulation methodologies - Markov-chain modelling, Agent-Based Modelling (ABM), and Social Digital Twins (SDTs) - that are fundamentally better suited to the complexity and uncertainty of social systems. The paradigm shifts from seeking a single, deterministic answer about an individual to exploring a range of possible systemic futures under different policy conditions.

4.1 Foundational Methodologies for System Simulation

These three methodologies represent a spectrum of complexity and capability, each offering a unique lens through which to analyse the children's services system.

Markov-Chain Modelling: At its core, a Markov chain is a stochastic model that describes the probability of an entity transitioning between a series of discrete states over time. The defining "Markov property" is that the probability of moving to the next state depends *only* on the current state, not on the history of how that state was reached.

This approach is highly effective for modelling a child's journey through the different legal and placement statuses within the care system. For example, a model could define states such as "at home with family," "kinship care," "foster care," "residential care," and "permanent placement (adoption)." By analysing historical data, one can calculate the transition probabilities between these states. This allows policymakers to visualize the entire system of pathways, identify bottlenecks (e.g., where children get "stuck" in temporary placements), and understand the long-term flow of the child population through the system.

Agent-Based Modelling (ABM): ABM is a bottom-up computational method that simulates a system as a collection of autonomous, interacting "agents". These agents are programmed with a set of attributes and behavioural rules, and they interact with each other and with a simulated environment.

ABM is uniquely suited for studying *emergent phenomena*: complex, system-level patterns that are not programmed into the agents but arise organically from their localized interactions. In a child welfare context, agents could represent children, parents, social workers, and even institutions like schools. By giving these agents rules based on established social science theories (e.g., how parental stress is affected by unemployment), a model can explore how a policy change, such as an increase in financial support for families, might ripple through the system, affecting family dynamics, social worker caseloads, and ultimately, the number of children entering care. It allows for the testing of counterfactual scenarios, "what if?" questions, that would be impossible or unethical to conduct in the real world.

Social Digital Twins (SDTs): An SDT represents the most advanced and ambitious application of simulation. It is a dynamic, virtual representation of a real-world societal system, which integrates multiple data sources in near-real time to maintain an accurate, evolving model of the physical system. Unlike a static model, an SDT has a two-way data flow: it is updated by real-world data, and changes made in the twin can be used to inform changes in the real world.

An SDT of a local authority's children's services department would function as a "digital sandbox" for policy and strategic planning. It would move beyond simple "what if?" scenarios to allow for the testing of complex, interacting interventions.

For instance, a leadership team could simulate the combined impact of investing £5 million in new family support hubs, a 10% reduction in social worker caseloads, and the closure of a major local employer. The SDT would model the likely effects of this complex scenario on system-level outcomes like referral rates, placement stability, and overall budget pressure, providing invaluable strategic foresight before committing real-world resources.

4.2 Blueprint for a Children's Services System Simulator

A practical and effective simulator for children's services would not rely on a single methodology but would integrate these approaches into a multi-scale architecture.

Micro-level (The Family Agent): At the core of the simulator would be an Agent-Based Model. Each child and family unit would be modelled as agents, endowed with a set of attributes (e.g., age, income, housing stability, social support network, parental health) and behavioural rules. These rules would not be arbitrary but would be grounded in established social science theories like the Family Stress Model, which links stressors to parental wellbeing and parenting quality, or the Theory of Planned Behaviour, which models decision-making processes. (Hu and Keller 2015)

Meso-level (The Service Ecosystem): The family agents would exist within a simulated environment representing the local service ecosystem. This would include other agents, such as social workers (with their own attributes like caseload and experience level), and static resources like schools, health clinics, and community support centres. The accessibility and quality of these services would directly influence the behaviour and wellbeing of the family agents.

Macro-level (The Policy Environment): High-level policy and funding decisions would act as the user-controlled parameters of the simulation. A policymaker could use the simulator's interface to change variables such as the overall departmental budget, the eligibility criteria for early help services, or the number of social workers employed. These macro-level changes would cascade down, altering the constraints and opportunities at the meso and micro levels. (Goldhaber-Fiebert et al. 2012)

Purpose: Policy Testing, Not Individual Prediction: It is crucial to reiterate the fundamental shift in purpose. The primary output of such a simulator would not be a "risk score" for an individual family. Instead, it would be a dashboard of system-level performance indicators: projected number of children in care, average social worker caseload, inequalities in outcomes between different demographic groups, and overall budget expenditure. The goal is not to predict a single future, but to compare the likely systemic consequences of competing policy strategies, enabling a more informed and evidence-based design process.

This shift from prediction to simulation represents a move from a deterministic to a probabilistic and exploratory worldview. A predictive model offers a single, falsely precise answer: "this child's risk score is 17." A simulation model, by contrast, offers a distribution of possible futures: "under this policy, the number of children entering care is likely to fall within this range, with these potential unintended consequences for other parts of the system."

Because ABMs are stochastic, running the same simulation multiple times will produce a range of results, reflecting the inherent unpredictability of the real world. This approach embraces uncertainty rather than masking it. It transforms technology from a tool that claims to provide

definitive answers into a tool that helps decision-makers ask better questions, understand systemic trade-offs, and design policies that are robust across a range of possible futures.

Section 5: Navigating the Moral Landscape of Social Simulation

The proposed shift from predictive profiling to systems-level simulation is motivated by a desire to escape the profound ethical harms of the former. Simulation avoids the injustice of assigning a deterministic risk score to an individual and instead focuses on evaluating the fairness and effectiveness of the system itself.

However, this new paradigm is not ethically neutral. It introduces a new set of complex systemic ethical challenges that demand a purpose-built governance framework. While simulation is a step forward, its responsible implementation requires a rigorous application of existing ethical principles and the development of new safeguards to address the unique risks posed by dynamic, self-learning digital twins.

5.1 Applying Existing Frameworks in a New Context

Any ethical framework for social simulation must be grounded in established global standards. The UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO 2021) offers a human-rights-centred foundation, while the UK's Data Ethics Framework (GDS 2020) and approach laid out in the previous government's AI White Paper (DSIT 2023) translate these principles into actionable guidance for public institutions. Together, these frameworks provide a dual anchor, universal values and jurisdiction-specific operational safeguards, that can be adapted for the context of a children's services simulator.

Human Oversight and Determination: This is the paramount principle. The simulator must be understood and governed as a decision-support tool, never a decision-maker. Its outputs are not directives to be followed blindly but are pieces of evidence to be critically examined, debated, and integrated with professional judgment, qualitative evidence, and democratic deliberation. Ultimate responsibility and accountability for policy decisions must always rest with human beings.

Transparency and Explainability: The "black box" problem that plagues many machine learning models is unacceptable in this context. The simulator's design must be transparent. Its core assumptions, the social science theories embedded in its agent rules (e.g., the Family Stress Model), and its data sources must be open to scrutiny by stakeholders. Its logic must be explainable, allowing users to understand *why* a particular policy scenario produced a given outcome.

Fairness and Non-Discrimination: A primary purpose of the simulator must be to identify and mitigate systemic biases, not to ignore or perpetuate them. It should be explicitly used to run

scenarios that assess the differential impact of proposed policies on various demographic groups (e.g., by race, ethnicity, socioeconomic status, and geography). This allows policymakers to proactively design interventions that advance equity rather than inadvertently deepening existing inequalities.

Multi-stakeholder and Adaptive Governance: The design, validation, and ongoing use of the simulator cannot be the exclusive domain of technologists and government officials. A continuous and inclusive governance process is essential, involving social workers, academic researchers, ethicists, legal experts, and, most importantly, individuals with lived experience of the care system: children, parents, and care leavers. Their insights are indispensable for ensuring the model reflects real-world complexity and aligns with societal values.

5.2 The Unique Challenges of Digital Twins and Feedback Loops

While the UNESCO principles provide a necessary foundation, the move towards advanced Social Digital Twins (SDTs) introduces novel ethical risks that require additional safeguards. These risks stem from the dynamic, two-way data flow between the model and the real world, which can create dangerous and self-reinforcing feedback loops.

The Risk of Reifying Flawed Assumptions: A simulator is not a purely objective reflection of reality; it is a model based on theories. These theories are often derived from historical data, which, as the APC problem demonstrates, is imbued with the confounded and potentially biased social patterns of the past. There is a significant risk that the simulator's rules could embed these historical biases. If the model is then treated as an authoritative "twin" of reality, these biased assumptions could become laundered and entrenched under a veneer of technological sophistication, making them even harder to challenge.

The Peril of Performative Feedback Loops: The most profound ethical risk of a true SDT is the creation of performative feedback loops that blur the line between observing reality and creating it. A standard simulation models a hypothetical "what if?" scenario. An SDT, continuously updated with real-time data and used to inform real-time decisions, can create a pernicious cycle:

1. The SDT, based on its model, simulates an emerging problem (e.g., rising family stress) in a specific neighbourhood.
2. Policymakers, trusting the model, react by redirecting resources and increasing service intensity in that neighbourhood.
3. The increased intervention and surveillance in that neighbourhood inevitably generate more administrative data (e.g., more referrals, more assessments, more child protection plans).
4. This new data is fed back into the SDT, which interprets it as a confirmation of its original "prediction" that the neighbourhood was high-risk. The model is no longer merely describing reality; its predictions are actively shaping that reality in a way that confirms

the prediction. This can lead to self-fulfilling prophecies, the entrenchment of spatial stigma, and the amplification of the very inequalities the system is supposed to combat.

To mitigate these unique and severe risks, a more specialized and dynamic governance framework is required. The five-pillar model proposed for self-learning digital twin architectures provides a powerful blueprint for this purpose.

Proposed Ethical Governance Framework for a Children's Services Digital Twin

Pillar	Core Principle	Practical Application in Children's Services
1. Ethical-by-Design Integration	Proactively embed ethical constraints into the system's core architecture.	The simulator's code must include "fairness-aware" algorithms that prevent optimization for one outcome (e.g., cost reduction) at the expense of equity. Privacy-preserving techniques must be built in from the start.
2. Continuous Assurance Monitoring	Implement real-time monitoring to detect model drift, anomalies, and emergent biases.	A live dashboard, accessible to the oversight board, must track the simulator's outputs and compare them against real-world data, flagging any significant deviations or signs that the model is disproportionately targeting certain groups.
3. Transparent Decision Traceability	Maintain a clear, interpretable, and auditable log of the model's processes and outputs.	Every simulation run must be logged, detailing the input parameters (the policy tested) and the full range of output metrics. The system

		must be able to explain <i>why</i> a given input led to a given output in a way that is understandable to non-technical stakeholders.
4. Multi-Stakeholder Oversight	Establish an independent, empowered, and diverse governance body to oversee the system's entire lifecycle.	A permanent, independent Ethics & Oversight Board must be established. Its membership must include care-experienced young people, parents, frontline social workers, data scientists, ethicists, and community leaders. This board must have the authority to audit the system and veto proposed policy simulations.
5. Dynamic Compliance Adaptation	Ensure the governance system can evolve in response to changes in the model, the law, and societal values.	The governance framework must not be static. It must be reviewed annually by the Oversight Board, with protocols in place to automatically update the model's constraints in line with new legislation (e.g., data protection laws) or evolving ethical standards.

Conclusion and Recommendations

The children's services system in the UK is failing its most vulnerable citizens. It is caught in a systemic trap of rising costs, reactive crisis management, and devastatingly poor outcomes. The temptation to seek a technological panacea is strong, but the history of such interventions is a

stark warning. The paradigm of centralized surveillance, embodied by ContactPoint, failed because it decontextualized data and undermined professional judgment.

The paradigm of predictive analytics, embodied by the Allegheny Family Screening Tool, is failing because it is built on a statistically invalid foundation, ignoring the unsolvable Age-Period-Cohort problem. This serves to launder and amplify deep-seated societal biases under a false veneer of scientific objectivity.

Continuing down this path is not only ineffective but also profoundly unjust.

A radical new direction is required. This paper has argued for a fundamental paradigm shift: away from the flawed and unethical goal of predicting individual futures, and towards the constructive and collaborative task of designing better systems. Simulation methodologies, from Markov chains and Agent-Based Models to Social Digital Twins, offer a powerful toolkit for this purpose. They allow policymakers to move beyond reactive management and engage in proactive design, testing the systemic consequences of their decisions in a safe, virtual environment. They are tools not for generating definitive answers, but for fostering a deeper understanding of complexity, exploring a range of possible futures, and asking better questions.

This new paradigm is not without its own significant ethical challenges, particularly the risk of creating self-fulfilling prophecies through performative feedback loops. Its responsible implementation is contingent upon the creation of a robust, transparent, and participatory governance framework that places human rights, professional judgment, and lived experience at its core.

Based on this analysis, the following recommendations are proposed:

For Policymakers and Government Departments:

Institute a Moratorium on Predictive Risk Modelling: Cease the procurement and deployment of predictive analytics tools for individual risk assessment in child welfare, pending a thorough review of their statistical validity and ethical implications in light of the Age-Period-Cohort problem.

Invest in Simulation Capabilities: Redirect funding from predictive analytics projects towards building internal and cross-sector capacity for systems-level simulation. This should include pilot projects using Agent-Based Modelling to test specific policy interventions at a local authority level.

Establish a Centre for Social Simulation: Create a multi-disciplinary body, modelled on the principles of Social Finance's Data + Digital Labs and Data to Insight, to develop open-source

simulation tools, establish best practices, and support local authorities in adopting these new methods.

For Technologists and Data Scientists:

Shift Focus from Prediction to Explanation: Reorient research and development away from creating "black box" predictive models and towards building transparent, explainable simulation models that illuminate the underlying mechanisms of social systems.

Embrace Co-Design: Adopt participatory design methodologies that involve social workers, families, and care-experienced young people throughout the entire lifecycle of model development, from defining the problem to validating the results.

Prioritize Ethical Governance: Champion the adoption of rigorous governance frameworks, such as the five-pillar model outlined in this report, as an integral and non-negotiable component of any social simulation project.

For Researchers and Academic Institutions:

Advance the Science of Social Simulation: Conduct further research on validating agent-based models in social policy contexts and develop new methods for integrating qualitative data and social theory into computational models.

Educate the Next Generation: Integrate modules on systems thinking, computational social science, and the ethics of AI and simulation into social work, public policy, and data science curricula to build a future workforce equipped with the necessary skills for this new paradigm.

The challenges facing children's services are immense, but they are not insurmountable. They are, at their core, problems of system design. By abandoning the false promise of prediction and embracing the exploratory power of simulation, we can move from a posture of managing failure to one of designing success, creating a system that is not only more efficient and effective, but also fundamentally fairer and more humane.

References

- Allegheny County. n.d. 'Allegheny Family Screening Tool'. Accessed 9 October 2025. <https://www.alleghenycounty.us/Services/Human-Services-DHS/DHS-News-and-Events/Accomplishments-and-Innovations/Allegheny-Family-Screening-Tool>
- Barnardo's. 2021. No Place Like Home - Parliamentary Briefing.
- Bell, Andrew, and Kelvyn Jones. 2015. 'Period and Cohort Processes in Longitudinal and Life Course Analysis: A Multilevel Perspective'. In A Life Course Perspective on Health Trajectories and Transitions. <https://www.ncbi.nlm.nih.gov/books/NBK385364/>
- CIPFA. 2025. 'Childrens Services'. CIPFA. <https://www.cipfa.org/policy-and-guidance/performance-tracker/childrens-services>
- DSIT. 2023. A Pro-Innovation Approach to AI Regulation. Dandy Booksellers Ltd.
- Eubanks, Virginia. 2018. Automating Inequality.
- GDS. 2020. Data Ethics Framework. https://assets.publishing.service.gov.uk/media/5f74a4958fa8f5188dad0e99/Data_Ethics_Framework_2020.pdf
- Glenn, Norwal D. 2005. Cohort Analysis. Vol. 5.
- Goldhaber-Fiebert, Jeremy D., Stephanie L. Bailey, Michael S. Hurlburt, et al. 2012. 'Evaluating Child Welfare Policies with Decision-Analytic Simulation Models.' Administration and Policy in Mental Health (United States) 39 (6): 466–77. <https://doi.org/10.1007/s10488-011-0370-z>
- Hu, Xiaolin, and Nicholas Keller. 2015. 'Agent-Based Modeling and Simulation of Child Maltreatment and Child Maltreatment Prevention'. Journal of Artificial Societies and Social Simulation 18 (3): 6. <https://doi.org/10.18564/jasss.2783>
- Samant, Anjana, Noam Shemtov, Kath Xu, et al. 2023. 'The Allegheny Family Screening Tool's Overestimation of Utility and Risk'. Logic(s) Magazine, no. 20 (December). <https://logicmag.io/policy/allegheny-family-screening-tools-overestimation-of-utility-and-risk/>
- TASO. 2025. 'Pathways into and through Higher Education for Young People with Experience of Children's Social Care'. TASO, March 6. <https://taso.org.uk/libraryitem/pathways-higher-education-for-young-people-childrens-social-care/>
- UKRI. 2025. 'Innovative Data Use Reveals Children's Care System Outcomes'. UKRI, March 3. <https://www.ukri.org/news/innovative-data-use-reveals-childrens-care-system-outcomes/>

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization (UNESCO).

<https://unesdoc.unesco.org/ark:/48223/pf0000380455>

Wikipedia. 2025. 'ContactPoint'. September 15.

<https://en.wikipedia.org/w/index.php?title=ContactPoint&oldid=1311423350>

- *Inteligencia artificial, ejecución penal y derechos humanos: ¿interacción inconciliable o posible cohesión?*

Federico M. Alessio Voskerichian. Abogado, Profesor en Derecho, Especialista en Derecho Penal (Universidad Nacional de Rosario). Maestrando en Derechos Fundamentales (Universidad de Granada).

- *Del algoritmo al sesgo cognitivo y de la hiperconexión a la soledad. Impacto de la IA y digitalización en los Derechos Humanos de menores y personas mayores*

Lucía Aparicio Chofré. Universitat de València.

- *Neocolonialismo algorítmico y captura regulatoria: desafíos jurídicos en la investigación farmacéutica global*

Itziar Artiñano Ortiz. PhD Candidate, UCM.

- *Regulación de la IA y riesgos democráticos: un crítico análisis de los derechos humanos en El Salvador*

Pablo Ceballos Borredà. Universitat de València.

- *Opacidad algorítmica y derechos humanos: comparación crítica UE-EEUU*

Carlos García-Llorente e Ignacio Olmeda. Universidad de Alcalá.

- *La inteligencia artificial: ¿una distopía para los derechos humanos?*

Nacho Hernández Moreno. Doctor en Derechos Humanos, Democracia y Justicia Internacional (Universitat de València).

- *Mecanismos de control para prevenir los sesgos en la herramienta VioGén*

Mireia Molina. Ayudante doctora. Universitat de València.

- *La IA como promotora de derechos: inclusión social y accesibilidad universal*

Mar Monsoriu Flor. Divulgadora tecnológica.

- *Delitos de odio y discurso violento en redes sociales*

Carla de Paredes Gallardo, Laura Lopez-Almansa Beaus y Susana Berrocal Díaz. Facultad de Ciencias Sociales. Universidad Europea de Valencia.

- *CIANDO por los derechos humanos*
Sergio Pérez González. Profesor Titular de Derecho Penal, Universidad de La Rioja (UR). **David San Martín Segura.** Profesor Permanente Laboral de Derecho Administrativo, UR.
- *Voluntary Carbon Credits, activos digitales e inteligencia artificial: hacia una gobernanza global basada en derechos*
Paula Sáez Álvarez. Profesora Permanente Laboral. Universidad CEU Fernando III, CEU Universities.
- *PROMETEO LIBERADO: la tecnología de la inteligencia artificial y el Derecho Administrativo*
Davi Augusto Santana de Lelis. Universidade Federal de Viçosa, Brasil.
- *Protección de los derechos humanos de las personas extranjeras en los procedimientos administrativos de extranjería: actuaciones automatizadas*
Paula María Serrano Casaus. Universitat Jaume I.
- *La inteligencia artificial y su impacto en los derechos humanos. Ámbito normativo y desafíos que se deben afrontar*
María Lourdes Soto Rodríguez. Jueza Sustituta Adscrita al Tribunal Superior de Justicia de Galicia.
- *Automatización algorítmica y Derechos Humanos: un llamamiento a una gobernanza ética global*
Mercedes Ten Doménech. Profesora de Filosofía del Derecho en la Universidad CEU Cardenal Herrera.
- *Los neuroderechos en el horizonte de la singularidad tecnológica*
Aitana Torró Calabuig. Universitat de València.
- *Retos y desafíos que plantea la Inteligencia artificial en las actuales democracias*
Carmen Ventura Salom. Licenciada en CC. Políticas y de la Administración. Doctora en Derecho (UV) y Máster en Estudios Internacionales y de la Unión Europea.

Lucía Aparicio Chofré. Del algoritmo al sesgo cognitivo y de la hiperconexión a la soledad. El impacto de la IA y la digitalización en los Derechos Humanos de los colectivos más vulnerables: menores y personas mayores

Lucía Aparicio Chofré

Departamento de Filosofía del Derecho y Política. Universitat de València

[ACCEDE A LA GRABACIÓN](#)

Resumen:

La extensión generalizada de la Inteligencia artificial y la digitalización masiva en nuestra vida cotidiana plantean una serie de retos, desafíos y posibles peligros para los colectivos más vulnerables. En especial los menores y las personas mayores, que se expondrán en las siguientes líneas, desde una perspectiva basada en Derechos Humanos y guiada por el principio de desarrollo sostenible prestando una especial atención al principio comunitario de prevención y a su pionera regulación a nivel comunitario.

En primer lugar, se presentarán una serie de datos estadísticos, de conformidad con el ODS 17, Meta: 17.8.10, Indicador: 17.8.111, sobre la extensión acelerada del empleo de esta nueva tecnología a nivel global.

Posteriormente, se analizará el impacto que, lejos de ser neutro, esta tecnología supone en materia de Derechos Humanos. En particular en relación con los derechos a la salud, igualdad, educación o ambiente; para ello se destacarán una serie de relevantes estudios científicos y

¹⁰ Poner en pleno funcionamiento, a más tardar en 2017, el banco de tecnología y el mecanismo de apoyo a la creación de capacidad en materia de ciencia, tecnología e innovación para los países menos adelantados y aumentar la utilización de tecnologías instrumentales, en particular la tecnología de la información y las comunicaciones

¹¹ Proporción de personas que usan Internet.

médicos que ya constatan y alertan de los serios peligros de su utilización, tanto a nivel físico como psicológico, en el bienestar y correcto desarrollo de estos colectivos.

Se concluirá resaltando la experiencia práctica, de sensibilización y resistencia, de la Asociación de la Adolescencia Libre de Móviles de la Comunidad Valencia, pionera a nivel nacional y de la que la autora forma parte y formulando, finalmente, una reflexión iusfilosófica con el objetivo de minimizar posibles riesgos y promover un desarrollo humano sostenible.

Palabras clave: Inteligencia Artificial, capacidad crítica, fuentes científicas relevantes, Objetivos de Desarrollo Sostenible, Agenda 2030 y Derechos Humanos.

Sumario: 1- La acelerada extensión de la Inteligencia Artificial y la digitalización masiva. 2. El impacto de la IA en los Derechos Humanos y los Objetivos de Desarrollo Sostenible, en especial en menores y personas mayores. 3. La Asociación de la Adolescencia Libre de Móviles de la Comunidad Valencia como experiencia pionera de sensibilización y resistencia. 4. Referencias bibliográficas.

1- La acelerada extensión de la Inteligencia Artificial y la digitalización masiva

En 2024, el 68 % de la población mundial tenían acceso a Internet es decir 5.500 millones de personas según el último informe de Naciones Unidas en relación con el progreso en los Objetivos de Desarrollo Sostenible publicado el pasado mes de julio¹². Sin embargo, todavía un tercio de la población mundial, a saber 2.600 millones de personas, no lo tiene. En este sentido, casi en la última década, la utilización mundial de Internet ha pasado de ser un 40 % de la población mundial en 2015 al 68 % en 2024. Un incremento que equivale a una tasa de crecimiento medio anual del 6,1 %, pero aun así, el objetivo de lograr un uso universal fijado en el Objetivo de Desarrollo Sostenible (ODS) nº 17 “Alianzas para lograr los Objetivos”¹³, con una penetración del 95 %, sigue siendo todavía una perspectiva lejana.

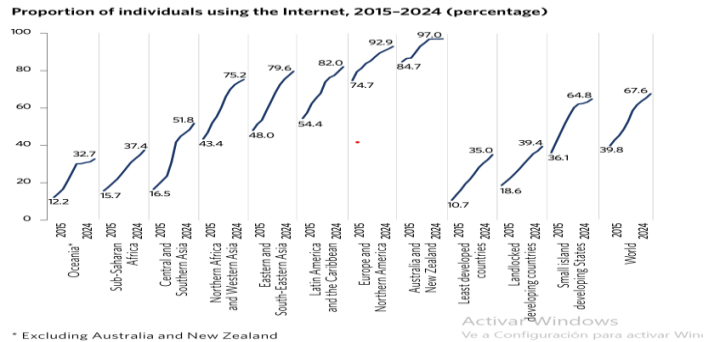
¹² Naciones Unidas, *The Sustainable Development Goals Report 2025*, 14 de julio de 2025, disponible en: <https://unstats.un.org/sdgs/report/2025>

¹³ En concreto el ODS 17 en sus metas relacionadas con la tecnología (ODS 17.6, 17.7 y 17.8) establece al respecto lo siguiente: 17.6 Mejorar la cooperación regional e internacional Norte-Sur, Sur-Sur y triangular en materia de ciencia, tecnología e innovación y el acceso a estas, y aumentar el intercambio de conocimientos en condiciones mutuamente convenidas, incluso mejorando la coordinación entre los mecanismos existentes, en particular a nivel de las Naciones Unidas, y mediante un mecanismo mundial de facilitación de la tecnología.

17.7 Promover el desarrollo de tecnologías ecológicamente racionales y su transferencia, divulgación y difusión a los países en desarrollo en condiciones favorables, incluso en condiciones concesionarias y preferenciales, según lo convenido de mutuo acuerdo.

17.8 Poner en pleno funcionamiento, a más tardar en 2017, el banco de tecnología y el mecanismo de apoyo a la creación de capacidad en materia de ciencia, tecnología e innovación para los países menos adelantados y aumentar la utilización de tecnologías instrumentales, en particular la tecnología de la información y las comunicaciones.

Así el uso de Internet sigue estando estrechamente vinculado al nivel de desarrollo. El uso universal se ha logrado o casi se ha logrado en Australia y Nueva Zelanda, así como en Europa y América del Norte, mientras que el África subsahariana y Oceanía (excluidas Australia y Nueva Zelanda) se están quedando atrás, con solo el 37 % y el 33% de la población conectada, respectivamente. Entre los países menos adelantados (PMA) y los países en desarrollo sin litoral (PDSL), el uso de Internet se sitúa en solo el 35 % y el 39 %, respectivamente.



A nivel mundial según el sexo, el 70 % de los hombres tenían conexión, frente al 65 % de las mujeres, lo que supone 189 millones más de hombres que de mujeres conectados en 2024. Una diferencia que como resalta Naciones Unidas ha ido disminuyendo desde 2021, cuando era de 277 millones.

A fin de alcanzar estas tres metas Naciones Unidas, en su Resolución aprobada por la Asamblea General, el 6 de julio de 2017, relativa a la Labor de la Comisión de Estadística en relación con la Agenda 2030 para el Desarrollo Sostenible (A/RES/71/313) estableció los siguientes cuatro indicadores, estando dos de ellos relacionados con el acceso a internet:

17.6.1 Número de acuerdos y programas de cooperación en materia de ciencia o tecnología suscritos por los países, desglosado por tipo de cooperación.

17.6.2 Número de abonados a Internet de banda ancha fija por cada 100 habitantes, desglosado por velocidad.

17.7.1 Total de los fondos aprobados para los países en desarrollo a fin de promover el desarrollo, la transferencia y la difusión de tecnologías ecológicamente racionales.

17.8.1 Proporción de personas que utilizan Internet.

Proportion of individuals using the Internet, by sex
(Percentage)

Regions	2019		2020		2022		2023		2024	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
World	50.4	55.5	55.9	61.8	60.8	66.5	63.0	67.7	65.3	69.5
Sub-Saharan Africa	20.3	29.3	22.8	32.6	27.0	38.8	29.0	40.7	31.3	45.5
Northern Africa and Western Asia	56.9	65.4	62.7	69.3	67.4	77.2	68.9	78.7	70.7	82.6
Northern Africa	54.9	60.2	62.9	65.5	63.9	73.3	64.6	73.6	...	...
Western Asia	58.7	69.8	62.5	72.6	70.7	80.6	72.9	83.2	...	...
Central and Southern Asia	26.4	35.6	34.7	48.5	40.9	52.1	43.4	53.1	46.1	56.8
Central Asia	61.5	63.4	68.0	70.0	77.5	84.1	81.1	87.4	...	...
Southern Asia	25.1	34.6	33.4	47.7	39.4	50.9	41.9	51.8	...	...
Eastern and South-Eastern Asia	62.1	64.6	67.7	69.0	74.3	76.4	77.5	77.5	80.8	79.1
Eastern Asia	66.1	68.6	71.5	71.6	75.8	77.0	79.2	77.4	...	...
South-Eastern Asia	52.2	54.6	58.4	62.5	70.8	74.9	73.4	77.7	...	...

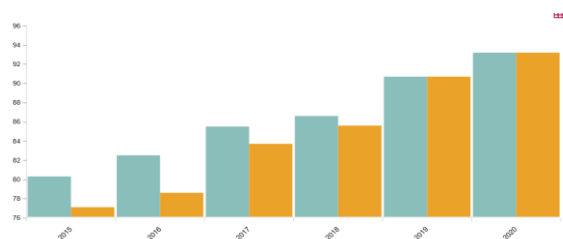
Regions	2019		2020		2022		2023		2024	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Latin America and the Caribbean	67.4	68.3	74.7	73.6	77.2	77.7	80.0	79.4	82.6	81.9
Oceania	73.4	74.4	75.5	76.8	76.9	78.4	77.0	78.4	...	...
Australia and New Zealand	93.0	92.9	94.8	94.8	96.3	97.5	96.4	97.5	96.5	99.3
Oceania (exc. Australia and New Zealand)	22.6	29.6	25.8	33.4	28.1	33.4	28.5	33.7	29.0	36.4
Europe and Northern America	84.0	86.3	87.2	90.3	90.4	90.6	91.5	91.7	92.7	93.2
Europe	81.4	84.5	84.5	86.9	88.7	90.5	90.1	91.5	...	...
Northern America	89.5	89.9	92.8	96.8	93.9	90.8	94.4	92.1	...	...
Landlocked developing countries	24.8	30.3	26.9	34.5	31.5	39.5	33.2	40.9	35.5	43.2
Least developed countries	18.6	25.2	20.7	29.5	24.9	36.3	26.6	38.3	28.7	41.1
Small island developing States	56.6	55.7	59.6	61.8	62.4	62.3	63.4	63.0	64.1	64.3

Source: Data based on the ITU DataHub, International Telecommunication Union (ITU).

En el caso de España, según el Instituto Nacional de Estadística¹⁴, el número de abonados a Internet de banda ancha fija por cada 100 habitantes, con una velocidad mayor o igual a 10 Mbit/s, pasó de 26,76 en 2015 a 34,92 en 2024.

Y la proporción de personas en edades comprendidas entre los 16 a 74 años que han utilizado Internet en los últimos tres meses pasó del 78.70% en 2015 al 95,40% en 2023.

Si se desglosan estos datos por sexos y edades, en relación con las cifras del año 2020, se puede apreciar como en estos últimos años se ha incrementado de forma significativa el número de mujeres hasta igualarse con el de hombres. Así, mientras en 2020, en los hombres era del 80,13% y en mujeres del 77,1% en 2020, en ambos casos se situaba en el 93,2%.



Por lo que respecta a las edades, aunque se excluyen a los menores de 16 años, en 2015 el porcentaje era del 98,5% en la franja de edad de 16 a 24 años, del 94,5% en las edades comprendidas entre los 25 y 34 años, del 91,6% entre los 35 y 44 años, del 82% entre los 45 y 54 años, del 61,4% entre los 55 y 64 años y finalmente del 31,3% entre los 65 y 74 años. En cambio, en el 2020 estos porcentajes fueron los siguientes, casi se igualó desde los 16 años a

¹⁴ Instituto Nacional de Estadística. Indicadores de la Agenda 2030 para el Desarrollo Sostenible, disponible en: <https://www.ine.es/dynqs/ODS/es/index.htm>

los 44 años situándose en el 99,8%, 99,7% y 99% respectivamente. También creció en los segmentos de los 45 a los 54 hasta situarse en el 97,1%, en el de los 55 a los 64 años hasta alcanzar el 89,5% y en de los 65 a los 74 años se duplicó con un 69,7%.

En cuanto al acceso a esta tecnología por parte de los menores se pueden destacar dos recientes estudios¹⁵. El uso de las tecnologías por los menores en España¹⁶ elaborado por el Observatorio Nacional de Tecnología y Sociedad en 2024 dependiente del Ministerio para la Transformación Digital y de la Función Pública y Móviles en España 2025: penetración, uso infantil y gestión sostenible¹⁷. Así de estos estudios se pueden extraer los siguientes datos:

El 70% de los niños entre 10 y 15 años poseen un dispositivo, cifra que alcanza el 96% en los adolescentes de 15 años y que se sitúa en el 22% a los 10 años. El porcentaje de niños con móvil crece significativamente a partir de los 10 años, y a los 12 años, más de dos tercios ya tienen uno (69%). Por sexo, aunque la diferencia es leve, las niñas (entre 10 y 15 años) tienen una mayor penetración de móviles, un 71%, en comparación con los niños, un 68%.

Según el tipo de hogar, los hogares más pequeños con niños (de 2 miembros) tienen la mayor penetración (75%), mientras que los hogares con 5 o más miembros (incluyendo infantes) presentan la menor (65%). Por otra parte, los hogares monoparentales con infantes presentan una mayor penetración de móviles (74%) en comparación con los hogares con una pareja y sus hijos (69%). También se puede destacar que no se aprecien diferencias significativas según el nivel de ingresos o educativo del hogar.

Por lo que se refiere a su uso se puede destacar como el 94,7% de la población de 10 a 15 años usó Internet en los últimos tres meses en 2023. El 93,1% utilizó ordenadores o tabletas. Las niñas (95,4%) usan más Internet que los niños (94%), con una diferencia de 1,4 puntos porcentuales. Respecto a 2022 se ha producido un ligero descenso de 2 décimas, mientras que la diferencia crece hasta los 2,8 puntos si se compara con 2021. No obstante, el uso de Internet en menores es superior al registrado en 2018, cuando lo usó el 92,8%.

¹⁵ Aunque también existen otros menos recientes como: Andrade, B., Guadix, I., Rial, A. y Suárez, F. *Impacto de la tecnología en la adolescencia. Relaciones, riesgos y oportunidades*, UNICEF España, Madrid, 2021, disponible en: https://www.unicef.es/sites/unicef.es/files/comunicacion/Informe_estatal_impacto-tecnologia-adolescencia.pdf

¹⁶ El uso de las tecnologías por los menores en España. Observatorio Nacional de Tecnología y Sociedad. Red.es. Secretaría de Estado de Digitalización e Inteligencia Artificial. Ministerio para la Transformación Digital y de la Función Pública. Reservados todos los derechos. Edición 2024 - Datos 2023 disponible en <https://www.ontsi.es/sites/ontsi/files/202407/El%20uso%20de%20menores%20en%20Espa%C3%B1a%202023.pdf>

¹⁷ Móviles en España 2025: penetración, uso infantil y gestión sostenible, TBS Education, Barcelona, 2025, disponible en <https://www.observatoriodelainfancia.es/oia/esp/descargar.aspx?id=8892&tipo=documento>

Además, el uso de Internet en menores crece progresivamente entre los 10 y los 15 años. El 91,7% de los niños y niñas de 10 años lo utilizan frente al 97,1% de quienes tienen 15 años. La mayor variación entre edades está en el paso de los 11 a los 12 años, con un incremento de 2,9 puntos porcentuales.

Unos datos que sin duda plasman el nivel de extensión y penetración de estas tecnologías, en especial entre los más jóvenes con las consecuencias que pueden conllevar y que a continuación se detallarán.

2. El impacto de la IA y la digitalización en los Derechos Humanos, en especial en menores y personas mayores

En relación con el impacto de esta nueva tecnología en los Derechos Humanos, en especial en relación con el Derecho a la Salud, pero también en el Derecho a la intimidad y a la Educación, cada vez es más abundante la literatura científica¹⁸ e incluso la propia Organización Mundial de la Salud que advierte de los riesgos y consecuencias perjudiciales que puede generar en algunos colectivos vulnerables como los menores¹⁹.

A este propósito se puede destacar un reciente²⁰ estudio publicado en la Revista de Desarrollo Humano y Capacidades, en el que basándose en los datos del Global Mind Project para examinar los efectos a nivel poblacional que tiene la posesión de un teléfono inteligente durante la infancia sobre la salud mental y el bienestar en la edad adulta temprana, se constata

¹⁸ Así de forma meramente ejemplificativa se pueden citar las siguientes recientes publicaciones: Goñi, E. "Inteligencia artificial y menores", en *Digitalización del servicio público de justicia e inteligencia artificial judicial* / coord. por Donado, A., Jiménez Muñoz, F.J; Calaza, S. (dir.), Yáñez, F. (dir.), Dykinson, Madrid, 2024, pp. 153-176; Febles, N. "Menores, derecho y algoritmo en la era de la inteligencia artificial", *La Ley Derecho de Familia: Revista jurídica sobre familia y menores*, Vol. 12, Nº. 46 (Monográfico: *Inteligencia Artificial y familia: un cálido encuentro en plena encrucijada*), 2025, pp. 47-56, Abadías, A. *Child Grooming: el embaucamiento de menores en la era del Metaverso y la Inteligencia Artificial*, Tecnos, Madrid, 2024; Berrocal, A.I, "La protección jurídica de los menores y adolescentes ante el reto de la inteligencia artificial", *La Ley Derecho de Familia: Revista jurídica sobre familia y menores*, Vol. 12, Nº. 46 (Monográfico: *Inteligencia Artificial y familia: un cálido encuentro en plena encrucijada*), 2025, pp. 4-32; Fernández Aller, C., Gutiérrez García, E., Puente, C. Águeda, Essinger, M. *Inteligencia artificial y derechos fundamentales. Menores, cuidado y vulnerabilidad digital*, Documentos de trabajo (Laboratorio de alternativas), Nº. 238, 2024; Pardo de Vera, M. "La incidencia de la Inteligencia Artificial en los derechos digitales de los menores. Regulación legal y problemática actual", *La protección de la infancia en entornos digitales* / coord. por Rodríguez Rodríguez, D.M; España MartíM. (pr.), Dykinson, Madrid, 2024, pp. 123-130, Rebollo, L. *Inteligencia artificial y derechos fundamentales*, Dykinson, Madrid, 2023; Brión, E. *Protección de datos personales, redes sociales y menores de edad: Nuevos retos normativos en tiempos de big data e inteligencia artificial*, Dykinson, Madrid, 2025; Cóceres, M.C., "La inteligencia artificial como herramienta para la promoción de los derechos humanos de las personas mayores", *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo*, Vol. 13, Nº. 2, 2025, pp. 103-124.

¹⁹ Wu, D., Dong, X., Liu D.y Li H., "How Early Digital Experience Shapes Young Brains During 0-12 Years: A Scoping Review", *Early Education and Development*, nº 7, 2023, Hidalgo, D. *Anestesiados. La humanidad bajo el imperio de la tecnología*, Libros de la Catarata, Madrid, 2021 y Artigas, M. *Cerebro digital. Así nos transforma la tecnología*, National Geographic España, 2025.

²⁰ Thiagarajan, T., Newson, J. y Swaminathan, S. "Protecting the Developing Mind in a Digital Age: A Global Policy Imperative", *Journal of Human Development and Capabilities*, nº 26:3, 2025, pp. 493-504.

como recibir un teléfono inteligente antes de los 13 años se asocia con peores resultados de salud mental en la edad adulta temprana, especialmente entre las mujeres, incluyendo pensamientos suicidas²¹, desconexión de la realidad, peor regulación emocional y disminución de la autoestima. Unas correlaciones que se determinan por varios factores, entre ellos el acceso a las redes sociales, el ciberacoso, la alteración del sueño y las malas relaciones familiares. Esta tendencia se observa de forma sistemática en todas las regiones del mundo, siendo mayor su magnitud en los países de habla inglesa²².

²¹ A este respecto se puede destacar en Estados Unidos la entidad Social Media Victims Law Center que brinda apoyo y asesoramiento legal e interpone demandas colectivas contra las plataformas ante los casos de suicidio y otros daños padecidos por menores en las redes sociales, para una mayor información: <https://socialmediavictims.org/> También el documental *Can't Look Away: The Case Against Social Media*, dirigido por Matthew O'Neill y Perri Peltz en 2025.

²² A nivel europeo se puede destacar el proyecto Project MIMic: A project to understand the role of media in boys' and girls' lives, 2020-2025 en el que participan más de 2.000 jóvenes de Bélgica, Francia y Eslovenia para más información puede consultarse: <https://www.projectmimic.eu/#hero>. Entre sus resultados se puede destacar como la narrativa del «éxito» y la «vida perfecta» está omnipresente tanto en los mensajes mediáticos de los modelos a seguir como en los de los adolescentes. Un primer estudio reveló que los adolescentes que se esfuerzan mucho por alcanzar la perfección ven más mensajes de este tipo (es decir, publicaciones en redes sociales de otras personas que muestran lo perfectas que son) y, a la inversa, que los adolescentes que ven más mensajes de este tipo son también los que experimentan una mayor presión por ser perfectos. Alrededor del 40% de los adolescentes de los tres países indicaron que se encontraban con frecuencia (muy) con contenidos positivos en las redes sociales, como imágenes de apariencias impecables o estilos de vida ideales. Al mismo tiempo, se detectó que hasta el 61 % de los adolescentes sienten la presión de ser perfectos en todo lo que hacen. De ellos, el 38 % afirmó sentir esta presión debido a las altas expectativas (percibidas) de su entorno, incluyendo la sociedad, sus compañeros y sus padres. Un segundo estudio demostró que los adolescentes que escuchan canciones de artistas que afirman que su éxito se debe a su esfuerzo no adoptan necesariamente esas creencias ni se identifican con ellas. Así parece que solo ciertos grupos de adolescentes imitan estas creencias sobre el éxito promovidas en la música. Aun así, esos mensajes son bastante populares, ya que el 12 % de los adolescentes tiene canciones favoritas con letras que contienen mensajes sobre el trabajo duro como medio necesario para alcanzar el éxito. En otro estudio, se exploró cómo los influencers, las celebridades y los deportistas en las redes sociales pueden inspirar a los adolescentes a convertirse en ciudadanos comprometidos, aumentando su confianza para participar en causas políticas y sociales. Para muchos adolescentes, ser políticamente o socialmente activos se ha convertido en otro ideal al que sienten que deben aspirar. También se detectó que una mayor presencia de contenido político compartido por sus modelos a seguir favoritos en las redes sociales se asociaba con un mayor sentimiento de empoderamiento para convertirse en ciudadanos activos entre los adolescentes. Esta implicación política por parte de sus modelos a seguir sigue siendo poco frecuente, ya que el 70 % de los adolescentes indicaron que su persona favorita nunca o rara vez hablaba de cuestiones sociales. Un cuarto estudio sobre el bienestar digital entre los jóvenes eslovenos reveló que uno de cada tres jóvenes cree que el uso del teléfono inteligente reduce su eficiencia en las tareas, su productividad y su rendimiento académico. Uno de cada cuatro jóvenes siente que tiene dificultades en el ámbito emocional cuando utiliza un teléfono inteligente y uno de cada seis cree que el uso del teléfono inteligente perturba su comunicación y sus relaciones con los amigos. Un quinto estudio reveló que el 86 % de los jóvenes tiene experiencias positivas con la comunicación en línea y prospera digitalmente, pero difieren en cómo desarrollan estas experiencias positivas con la comunicación en línea. Un grupo de jóvenes (51 %) percibía un alto desarrollo digital y seguía desarrollándose con el tiempo. Otro grupo (49 %) percibía un menor desarrollo digital y su sensación de control sobre su comunicación en línea disminuyó con el tiempo. Los jóvenes con un mayor desarrollo digital eran más propensos a tener padres con habilidades digitales y padres que hablaban con sus hijos sobre el contenido en línea y tenían en cuenta la opinión de los niños en estas conversaciones.

En 2016, la Academia Americana de Pediatría alertó por primera vez del impacto del mundo digital sobre la salud, y en estos últimos años han ido aumentando progresivamente los ensayos clínicos que corroboran este vínculo²³.

Por lo que respecta a nuestro país, la Asociación Española de Pediatría (AEP)²⁴, a través del grupo de trabajo de Salud Digital del Comité de Promoción de la Salud, lanzó en 2023 el Plan Digital Familiar de la AEP, una guía de recomendaciones adaptada a las necesidades de cada familia y a la edad de los menores que la componen. Asimismo, elaboró otro documento de sugerencias titulado 'Impacto de los dispositivos digitales en la enseñanza'.

A este respecto la AEP ha alertado en distintos comunicados como el impacto del uso excesivo de pantallas en la infancia y adolescencia es multifactorial, afectando a diversas áreas relacionadas con la salud y el bienestar, que reducen la calidad de vida como el sueño²⁵,

²³ Santos, R. M. S., C. G. Mendes, G. Y. Sen Bressani, S. de Alcantara Ventura, Y. J. de Almeida Nogueira, D. M. de Miranda, and M. A. Romano-Silva, "The Associations between Screen Time and Mental Health in Adolescents: A Systematic Review." *BMC Psychology* 11 (1), 2023, p. 127, Sapien Labs, "The Youth Mind: Rising Aggression and Anger." *Rapid Report*, 2025, disponible en: <https://doi.org/10.17605/OSF.IO/3WFJX>, Costello, N., R. Sutton, M. Jones, M. Almassian, A. Raffoul, O. Ojumu, M. Salvia, M. Santoso, J. R. Kavanaugh, and S. B. Austin. 2023. "Algorithms, Addiction, and Adolescent Mental Health: An Interdisciplinary Study to Inform State-Level Policy Action to Protect Youth from the Dangers of Social Media." *American Journal of Law & Medicine*, nº 49 (2-3), 2023, pp. 135–172, Rosič, J., Schreurs, L., Janicke-Bowles, S., Vandenbosch, L. "Trajectories of digital flourishing in adolescence: The predictive roles of developmental changes and digital divide factors", *Child Development*, Volume 95, Issue 5, 2024, Vanhoffelen, G., Gonzalez, A., Schreurs, L., Giraudeau, C., & Vandenbosch, L. (2025). "The Perfect Li(f)e: A Longitudinal Study on Positive Social Media Content and European Adolescents' Perfectionism". *Communication Research*, 0(0), 2025.

²⁴ La Asociación Española de Pediatría es una sociedad científica que representa a cerca de 13.000 pediatras que trabajan tanto en el ámbito hospitalario como de atención primaria. Está integrada por todas las sociedades científicas de pediatría regionales y las sociedades de pediatría de las distintas especialidades. El principal objetivo de la asociación es velar por la adecuada atención sanitaria, fomentar el desarrollo de la especialidad, tanto en sus aspectos asistenciales como en los docentes y de investigación. Además de asesorar a todas aquellas instituciones competentes en asuntos que puedan afectar o afecten a la salud, desarrollo e integridad del niño y del adolescente, así como divulgar e informar a la población sobre cuestiones de salud infantil.

²⁵ Un mayor tiempo de uso de las redes sociales y videojuegos se asocia con permanecer un menor número de horas en la cama, acostarse más tarde y mayor latencia del inicio del sueño. Además, el uso de pantallas antes de acostarse produce: un aumento de la somnolencia diurna; una disminución de la somnolencia nocturna; una reducción de la secreción de melatonina; un retraso del reloj circadiano; y una disminución y retraso del sueño REM. Todo ello favorece el estado de ánimo depresivo, las alteraciones de la conducta, la disminución de la autoestima, y la alteración del desarrollo cerebral.

alimentación y nutrición²⁶, actividad física²⁷, riesgo cardiovascular²⁸, fatiga visual²⁹ y volumen cerebral³⁰.

Y destacan como los últimos estudios confirman distintos impactos en el neurodesarrollo a diferentes edades. De forma que el uso de un teléfono para premiar o distraer a niños de 1 a 4 años provoca que los niños exijan los dispositivos para calmarse y se frustren si se les niega. El uso rutinario de dispositivos para distraer o calmar genera dificultad para el desarrollo de estrategias de autogestión; ocasiona dependencia de las pantallas para la regulación de las emociones y determinará dificultades de autorregulación en etapas posteriores.

Además, la AEP en su última recomendación de 5 de diciembre de 2024 aumenta el rango de edad por debajo del que se considera que no debe exponerse a los niños a pantallas, que pasa de los 2 a los 6 años considerándose que no existe un tiempo seguro. Como excepción, y bajo supervisión del adulto, se pueden usar las pantallas para el contacto social en tiempos cortos y con un objetivo concreto, por ejemplo, que la persona que está al otro lado de la pantalla le cuente un cuento o le cante una canción.

Y subraya como en la adolescencia finaliza tanto la maduración del sistema límbico como la maduración de la corteza cerebral. Los medios digitales interfieren a esta edad de dos formas distintas: aumentando la activación de la región límbica, al estar expuestos a sistemas de gratificación inmediata, y disminuyendo la actividad frontal por efecto desplazamiento de los estímulos adecuados para la edad.

Además, la multitarea relacionada con las pantallas se asocia con peores resultados cognitivos, una disminución de la capacidad de filtrar las distracciones y un aumento de la impulsividad y disminución de la memoria de trabajo. De este modo, los adolescentes que pasan demasiado

²⁶ Algunos estudios asocian el tiempo frente al televisor con una dieta menos saludable, favoreciendo el consumo de alimentos hipercalóricos y, por consiguiente, aumentando el riesgo de sobrepeso y obesidad.

²⁷ Se ha demostrado que los adolescentes que ya tienden al sedentarismo pasan más tiempo usando pantallas. Por el contrario, la limitación de su uso favorece la práctica de actividad física. Así la AEP recomienda la realización de actividad física moderada o vigorosa durante un mínimo de 60 minutos diarios en niños y adolescentes.

²⁸ La falta de ejercicio físico por dedicar demasiado tiempo a las pantallas aumenta el riesgo de desarrollar síndrome metabólico. Por el contrario, se ha vinculado la reducción del tiempo de pantallas con una reducción de la presión arterial y un aumento del colesterol “bueno” (HDL).

²⁹ El ojo seco, la picazón ocular, el lagrimeo, la sensación de tener un cuerpo extraño en el ojo o la visión borrosa son algunos de los síntomas causados por el exceso de pantallas, así como el dolor céntrico lumbar, la fatiga general y la cefalea. Además, en la infancia y adolescencia favorece especialmente la miopía progresiva y el estrabismo agudo.

³⁰ La evidencia científica apunta a una disminución del espesor de la corteza cerebral en varias regiones, debido al uso excesivo de pantallas.

tiempo frente a una pantalla tienen más probabilidades de presentar dificultades cognitivas graves.

A la luz de estas evidencias la AEP recomienda de 7 a 12 años: la limitación a menos de una hora (incluyendo el tiempo escolar y los deberes) su utilización, limitar el uso de los dispositivos con acceso a Internet, priorizar los factores protectores: actividades deportivas, relaciones con iguales cara a cara, contacto con la naturaleza, sueño, alimentación saludable, etc. En caso que se decida la utilización de un dispositivo es recomendable: que sea bajo la supervisión de un adulto, con dispositivos fijos y evitar el baño y dormitorio, así como pactar límites claros previamente tanto en tiempo como en contenidos adaptados a la edad.

Para el caso de los menores entre 13 y 16 años la AEP recomienda: su utilización menos de dos horas (incluyendo el tiempo escolar y los deberes). En el caso de que se permita el acceso a dispositivos -sin ser la única medida que se tome- instalar herramientas de control parental, priorizar el uso de teléfonos sin acceso a internet y retrasar la edad del primer móvil inteligente (con conexión a internet).

Por otra parte la AEP recalca la necesidad urgente de adoptar medidas para luchar contra el uso excesivo de pantallas a las administraciones, tanto a los gobiernos como al sistema educativo, porque señala la función de la familia en la protección del impacto de las pantallas en sus hijos es importante, sin embargo, sería una irresponsabilidad como medida de protección a la infancia que toda la responsabilidad recaiga sobre estas³¹ e incluso considera que sería pertinente declararlo un problema de salud pública. De lo contrario, “los menores de colectivos vulnerables serán los que tengan mayor afectación en la salud y el desarrollo”.

En relación con las medidas adoptadas, a nivel tanto nacional como internacional y europeo, se puede destacar como el Gobierno de España ha desarrollado medidas relacionadas con la capacitación y la protección de las y los menores en el entorno digital.

En relación con la formación, se han planteado iniciativas orientadas a educar digitalmente a la ciudadanía y garantizar la inclusión digital, priorizando la provisión de competencias digitales a determinados colectivos entre los que se encuentran los niños, niñas y adolescentes en riesgo

³¹ A este respecto se argumentan dos razones fundamentales: el tiempo que los niños pasan en otros ámbitos como el sistema educativo y que hay familias que, por diversas circunstancias previas o sobrevenidas, carecen de la capacidad de llevar a cabo esta función.

de exclusión digital. Respecto a la protección, se han desarrollado iniciativas como el Programa CONFÍA de INCIBE³².

También se puede destacar la adopción de la Carta de Derechos Digitales, publicada en julio de 2021³³, en la que se reconoce la importancia de la protección de las personas menores de edad en el entorno digital, con un apartado dedicado exclusivamente a esta cuestión³⁴ y también a la brecha en el acceso digital por lo que respecta a los mayores³⁵.

En enero de 2024 el Gobierno de España aprobó la creación de un comité de personas expertas, orientado a diseñar una estrategia nacional para fomentar un entorno digital seguro para la

³² Este programa incluye ejes relacionados con el desarrollo de acciones orientadas, entre otros, a menores y que pasan por la concienciación y comunicación, la capacitación en ciberseguridad, la cooperación y coordinación para la consolidación de una cultura de ciberseguridad y el despliegue de herramientas y soluciones de ciberseguridad.

³³ Carta de derechos digitales disponible en: <https://espanadigital.gob.es/lineas-de-actuacion/carta-de-derechos-digitales>

³⁴ En su capítulo segundo relativo a Derechos de Igualdad en su apartado X titulado: “Protección de las personas menores de edad en el entorno digital” (p.12) se establecen los siguientes seis apartados:

1. Con arreglo a las potestades que les son propias y de acuerdo con la legislación aplicable, las personas progenitoras, tutoras, curadoras, representantes legales o personas que presten apoyo para el ejercicio de la capacidad jurídica, velarán por que las personas menores de edad hagan un uso equilibrado y responsable de los entornos digitales a fin de garantizar el adecuado desarrollo de su personalidad y de preservar su dignidad y sus derechos fundamentales.

2. Los centros educativos, las Administraciones y cualesquiera personas físicas o jurídicas que desarrollen actividades en entornos digitales en las que participen personas menores de edad están obligados, por la legislación aplicable, a la protección del interés superior de la persona menor y sus derechos fundamentales, especialmente los derechos a la intimidad, al honor y a la propia imagen, al secreto de las comunicaciones y el derecho a la protección de datos personales. Deberá recabarse su consentimiento, si es mayor de 14 años, o el de sus representantes legales, para la publicación o difusión de sus datos personales o su imagen a través de servicios de redes sociales.

Asimismo, se promoverá la implantación de procedimientos para la verificación de la edad, el derecho a recibir formación e información adecuada y adaptada a sus necesidades sobre los entornos digitales a los que accedan y el acceso a medios para solicitar y en su caso obtener la tutela de sus derechos frente a comportamientos o acciones lesivas o ilícitas.

3. Salvo en las excepciones previstas en las leyes, están prohibidos los tratamientos de la información de personas menores orientados a establecer perfiles de personalidad en entornos digitales. Ninguna práctica de perfilado podrá dirigirse a manipular o perturbar la voluntad de personas menores, incluido el perfilado con fines publicitarios.

4. Con arreglo a la normativa aplicable, en los entornos digitales las personas menores tendrán derecho a recibir información suficiente y necesaria sobre el uso responsable y adecuado de las tecnologías.

5. Las personas menores de edad pueden expresar libremente sus opiniones e ideas a través de medios tecnológicos, así como participar y expresar su opinión en los asuntos públicos que les afectan, conforme a los derechos que les son inherentes, incluyendo la libertad de pensamiento, de conciencia, de religión, de asociación y de celebrar reuniones con fines pacíficos en el entorno digital. Se potenciará el uso de las tecnologías para el pleno desarrollo de este derecho.

6. Se impulsará el estudio del impacto en el desarrollo de la personalidad de personas menores derivado del acceso a entornos digitales, así como a contenidos nocivos o peligrosos. Dicho estudio prestará particular atención a sus efectos en la educación afectivo-sexual, las conductas dependientes, la igualdad, la orientación sexual e identidad de género, así como a los comportamientos antidemocráticos, racistas, xenófobos, capacitistas, machistas, discriminatorios o propios del discurso del odio.

³⁵ Así en el apartado XII, siempre en el capítulo segundo relativo al Derecho a la Igualdad, bajo el título “Brechas de acceso al entorno digital” establece los siguientes dos apartados:

1. Se fomentará y facilitará el acceso de todos los colectivos a los entornos digitales y su uso y la capacitación para el mismo.

2. Se promoverán políticas públicas específicas dirigidas a abordar las brechas de acceso atendiendo a posibles sesgos discriminatorios basados en las diferencias existentes por franjas de edad, nivel de autonomía, grado de capacitación digital o cualquier otra circunstancia personal o social para garantizar la plena ciudadanía digital y participación en los asuntos públicos de todos los colectivos en mayor riesgo de exclusión social, en particular el de personas mayores, así como la utilización del entorno digital en los procesos de envejecimiento activo.

población infantil y adolescente³⁶, que en diciembre de 2024 presentó un informe en el que se proponían 107 medidas para crear entornos digitales seguros³⁷.

Por otro lado, el 25 de marzo de este mismo año, se remitió por el Gobierno a las Cortes Anteproyecto de Ley Orgánica para la protección de las personas menores de edad en los entornos digitales³⁸, que se tramitará de urgencia y en estos momentos se encuentra en la Comisión de Justicia del Congreso de los Diputados en la fase de presentación y debate de enmiendas³⁹.

A nivel europeo se puede destacar la Estrategia para un Internet Mejor para los Niños (Estrategia BIK+)⁴⁰, adoptada el 11 de mayo de 2022, mediante la que se pretende garantizar que los niños sean protegidos, respetados y empoderados en línea en la nueva Década Digital, en consonancia con los Principios Digitales Europeos y en la que se toman en consideración la resolución del Parlamento Europeo sobre los derechos de los niños⁴¹, las conclusiones del Consejo sobre la alfabetización mediática⁴² y la recomendación del Consejo, estableciendo una Garantía Infantil Europea⁴³.

³⁶ Los fundamentos de este acuerdo son tres: la necesidad de aprobar una legislación integral para la protección de menores en Internet, la necesidad de impulsar una estrategia de enfoque multidisciplinario desde la educación y la promoción del desarrollo de soluciones tecnológicas para impedir el acceso de menores a contenidos para adultos.

³⁷ Informe del Comité de personas expertas para el desarrollo de un entorno digital seguro para la juventud y la infancia, diciembre 2024, disponible en: <https://www.juventudeinfancia.gob.es/sites/default/files/noticias/Informe%20del%20comit%C3%A9%20de%20personas%20expertas%20para%20el%20desarrollo%20de%20un%20entorno%20digital%20seguro%20para%20la%20juventud%20y%20la%20infancia.pdf>

³⁸ Un anteproyecto que puede consultarse en el siguiente enlace: <https://www.mpr.gob.es/servicios/participacion/audienciapublica/Documents/VSGT%202024/2024-0921%20APLO%20menores%20entornos%20digitales/ANTEPROYECTO%20DE%20LEY%20ORG%C3%81NICA%20PARA%20LA%20PROTECCI%C3%93N%20DE%20LAS%20PERSONAS%20MENORES%20DE%20EDAD%20EN%20LOS%20ENTORNOS%20DIGITALES.pdf>

³⁹ Un iter de tramitación parlamentaria que se puede consultar en el siguiente enlace: https://www.congreso.es/es/proyectos-de-ley?p_p=id=iniciativas&p_p=lifecycle=0&p_p=state=normal&p_p=mode=view&iniciativas=mostrarDetalle&iniciativas_legislatura=XV&iniciativas_id=121/000052

⁴⁰ Una estrategia que se encuentra disponible en: <https://digital-strategy.ec.europa.eu/es/policies/strategy-better-internet-kids>

⁴¹ Resolución del Parlamento Europeo, de 26 de noviembre de 2019, sobre los derechos del niño con ocasión de la celebración del 30.º aniversario de la Convención de las Naciones Unidas sobre los Derechos del Niño. (2019/2876(RSP)) DO C 232 de 16.6.2021, pp. 2–11, disponible en: <https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=CELEX%3A52019IP0066>

⁴² Conclusiones del Consejo sobre la alfabetización mediática en un mundo en constante transformación 2020/C 193/06, ST/8274/2020/INIT, DO C 193 de 9.6.2020, pp. 23–28, disponible en: <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A52020XG0609%2804%29>

⁴³ Recomendación del Consejo por la que se establece una Garantía Infantil Europea, Bruselas, 4 de junio de 2021 (OR. en) 9106/21, disponible en: <https://data.consilium.europa.eu/doc/document/ST-9106-2021-INIT/es/pdf>

También la Declaración Europea sobre los Derechos y Principios Digitales de 2023 contempla la importancia de la protección y el empoderamiento de la infancia y juventud en el entorno digital⁴⁴.

Por su parte, el 29 de abril de 2024, el Consejo adoptó un Reglamento para prorrogar una medida de protección contra el abuso sexual de menores hasta el 3 de abril de 2026⁴⁵.

Más recientemente, el Informe sobre el estado de la Década Digital 2030 de 2025⁴⁶, señala como uno de los retos sociales urgentes relacionados con la digitalización en la UE que si bien la tecnología avanza, la transformación digital ha intensificado las vulnerabilidades y las desigualdades, afectando especialmente a los menores y a la salud mental. Una de las principales preocupaciones es la integridad de la información: el 88 % de los europeos se muestra preocupado por las noticias falsas y la manipulación en línea, y el 90 % considera que la protección de los niños en Internet es una prioridad fundamental. Estos riesgos, amplificados por la inteligencia artificial y las plataformas en línea, amenazan con socavar la integridad democrática, profundizar la polarización social y erosionar la confianza pública.

A nivel internacional, se puede destacar en el ámbito de la OCDE, la Recomendación del Consejo sobre los Niños en el Entorno Digital, aprobada en 2012 y enmendada en 2021⁴⁷ y, en mayo de 2025, publicó un informe titulado ¿Cómo es la vida de los niños en la era digital? (How's Life for Children in the Digital Age?)⁴⁸ en el que se ofrece una visión general del estado

⁴⁴ A tal fin, en su capítulo V, en sus apartados 20 a 22, se plantean una serie de compromisos entre los que se encuentran la capacitación digital, la promoción de experiencias digitales positivas en un entorno seguro y adaptado a su edad, la protección frente a potenciales amenazas o la realización de seguimientos y elaboración de perfiles y la implicación de los menores en el desarrollo de políticas digitales que les afecten. Declaración Europea sobre los Derechos y Principios Digitales para la Década Digital (2023/C 23/01), disponible en: <https://digital-strategy.ec.europa.eu/es/library/european-declaration-digital-rights-and-principles>

⁴⁵ En este reglamento se mantiene una excepción a las normas de protección de datos en el sector de las comunicaciones electrónicas que permite a los proveedores de los llamados servicios de comunicaciones interpersonales independientes de la numeración (por ejemplo, los servicios de mensajería) utilizar tecnologías específicas para el tratamiento de datos personales y de otro tipo con el fin de detectar abusos sexuales de menores en línea en sus servicios, denunciarlo y eliminarlo. Para un mayor detalle en esta materia puede consultarse el siguiente enlace: <https://www.consilium.europa.eu/es/policies/prevent-child-sexual-abuse-online/>

⁴⁶ Comisión Europea, State of the Digital Decade 2025: Keep building the EU's sovereignty and digital future, COM(2025) 290 final, Bruselas, 16 de junio de 2025, disponible en: <https://digital-strategy.ec.europa.eu/en/library/state-digital-decade-2025-report>

⁴⁷ En la que se reconocen una serie de principios para un entorno digital seguro y beneficioso para la infancia y su adaptación en el marco de políticas públicas de cada país a través de la aplicación de estas cinco recomendaciones: demostrar el liderazgo y compromiso con el interés superior del menor en el entorno digital; revisar, desarrollar y reformar aquellas leyes que afecten a los menores en línea; promover la alfabetización digital; adoptar políticas basadas en la evidencia para apoyar a los niños y niñas en el espacio digital y promover la adopción de la seguridad a través del diseño de servicios digitales adaptados a los menores. OCDE, Recommendation of the Council on Children in the Digital Environment, 2012, disponible en: <https://legalinstruments.oecd.org/public/doc/272/272.en.pdf>

⁴⁸ OCDE, How's Life for Children in the Digital Age?, 15 de mayo de 2025, disponible en: https://www.oecd.org/en/publications/how-s-life-for-children-in-the-digital-age_0854b900-en/full-report.html

actual de la vida digital de los niños en los países de la OCDE. Además en el que se incluyen 45 indicadores comparables a nivel internacional relacionados con el acceso y el uso de las tecnologías digitales por parte de los niños, las interacciones sociales en línea, las experiencias negativas y las prácticas que aumentan o reducen el riesgo. Así, por ejemplo, mientras que el 51 % de los estudiantes de 15 años de los países de la OCDE afirman ser capaces de ajustar fácilmente la configuración digital para proteger su privacidad, el 27,6 % afirman compartir información inventada en las redes sociales y el 17 % afirman sentirse nerviosos o ansiosos sin sus dispositivos digitales. Los datos también muestran que el ciberacoso está aumentando en todos los países de la OCDE y que una minoría significativa de niños informa de un uso problemático de las redes sociales. Además, se ha observado que pasar mucho tiempo conectado a Internet desplaza otras actividades valiosas y enriquecedoras, como la lectura, el ejercicio físico o la socialización presencial con amigos, que se consideran esenciales para el desarrollo cognitivo y socioemocional de los niños.

Por su parte en el ámbito de Naciones Unidas, el 22 de septiembre de 2024, la Asamblea General adoptó el denominado Pacto para el Futuro⁴⁹ en el que en su primer anexo se contiene un Pacto Digital Global.

También el Comité de Derechos del Niño, el 2 de marzo de 2021, adoptó la Observación general núm. 25 (2021) relativa a los derechos de los niños en relación con el entorno digital

⁴⁹ Un Pacto en el que, dentro de su primer objetivo a saber: Fomentar un espacio digital inclusivo, abierto y seguro que respete, proteja y promueva los derechos humanos, en materia de Confianza y seguridad digitales, a fin de contrarrestar y abordar urgentemente todas las formas de violencia, incluida la violencia sexual y de género, que se producen o amplifican por el uso de la tecnología, todas las formas de discurso de odio y discriminación, las informaciones erróneas y la desinformación, el ciberacoso y la explotación y los abusos sexuales de menores. Estableceremos y mantendremos estrictas medidas de mitigación de riesgos y vías de recurso que también protejan la privacidad y la libertad. Los Estados se comprometen de aquí al 2030 a crear un espacio digital seguro para todos los usuarios que garantice su salud mental y su bienestar definiendo y adoptando normas, directrices y acciones sectoriales comunes que respeten el derecho internacional, promuevan espacios cívicos seguros y aborden el contenido de las plataformas digitales que causa daño a las personas, teniendo en cuenta la labor que están realizando las entidades de las Naciones Unidas, las organizaciones regionales y las iniciativas multipartitas interesadas (ODS 3, 5, 9, 10, 16 y 17); Dar prioridad a la formulación y aplicación de políticas y normas nacionales de seguridad infantil en el entorno digital, de conformidad con el derecho de expresión internacional de los derechos humanos, en particular la Convención sobre los Derechos del Niño (ODS 3, 5 y 10); Y Vigilar y examinar las políticas y prácticas de las plataformas digitales sobre la lucha contra la explotación y los abusos sexuales infantiles que se producen o amplifican por el uso de la tecnología, incluida la distribución a través de plataformas digitales de abusos sexuales infantiles o material sobre abusos sexuales infantiles, así como la captación o seducción (grooming) de menores con el propósito de cometer delitos sexuales contra ellos (ODS 3). Resolución aprobada por la Asamblea General el 22 de septiembre de 2024 (A/RES/79/1) disponible en: <https://docs.un.org/es/A/RES/79/1>

(CRC/C/GC/25)⁵⁰. Además de otras iniciativas desplegadas por la Unión Internacional de Telecomunicaciones, UNICEF, UNESCO, etc⁵¹.

Por otra parte, la Unión Internacional de las Telecomunicaciones publicó, en 2020, las Directrices para la protección en línea de los niños⁵², con el fin de proveer orientaciones de actuación dirigidas a menores, padres y educadores, así como a empresas y decisores políticos ante los riesgos y amenazas de la digitalización sobre las y los menores. Y ha puesto en marcha un Programa global sobre protección infantil en línea y seguridad en Internet con Sango for Kids⁵³.

Finalmente, también el Foro Económico Mundial ha tratado la seguridad digital de los menores en un documento denominado Principios globales sobre seguridad digital: trasladando los derechos humanos internacionales al contexto digital, donde se menciona a la infancia como grupo vulnerable ante problemas como el abuso sexual o las interferencias en la privacidad.

A la luz de todas estas medidas cabría preguntarse su adecuación y pertinencia, para paliar los efectos nocivos que han sido ampliamente demostrados y si tal vez no sería conveniente ir un paso más allá, partiendo del consolidado principio europeo de precaución y adoptar medidas más estrictas como prohibir el acceso de los menores, como ha realizado Australia⁵⁴, a las redes sociales al igual que se hace con otro tipo de sustancias como el tabaco, el alcohol o actividades como el juego, que se consideran nocivas para la salud. A fin de retomar el control sobre nuestro futuro digital⁵⁵. De esta forma sería conveniente trasladar el foco de intervención de la responsabilidad de las familias y los menores a la de las plataformas y administraciones.

⁵⁰ En la que recomendaba a los Estados que tomaran medidas fuertes, entre ellas, la creación de legislación que proteja a los menores frente a contenidos nocivos y engañosos. Los niños también deben gozar de protección contra todas las formas de violencia presentes en el entorno digital, como la trata de niñas y niños, la violencia de género, la ciberagresión, los ciberataques y la guerra informática. Unas recomendaciones que han sido objeto de crítica por parte de la doctrina por su imprecisión, falta de concreción de responsabilidades, la dificultad de aplicación práctica o enfoque proteccionista. Véase entre otros: Bellver V. “Los derechos del niño en el entorno digital”, en Ballesteros, A. (ed.), *La digitocracia a debate*, Aranzadi, Pamplona, 2022, pp. 91-127.

Una Observación que se puede consultar en el siguiente enlace: <https://docs.un.org/es/CRC/C/GC/25>

⁵¹ En este sentido se pueden destacar las siguientes iniciativas: En materia de ciberseguridad la Iniciativa de Protección Infantil en línea una red compuesta por todas las partes interesadas lanzada por la Unión Internacional de Telecomunicaciones (UIT) para concienciar sobre la seguridad infantil en el mundo en línea y desarrollar herramientas para ayudar a los gobiernos, la industria y los educadores, para más detalle se puede consultar: <https://www.itu.int/en/ITU-D/Cybersecurity/Pages/COP/COP.aspx>, en materia de ciberacoso destaca la iniciativa Kindly de UNICEF, para más detalle véase: <https://www.unicef.org/innovation/kindly> o los proyectos Global Kids Online y Disrupting Harm.

⁵² Estas directrices pueden consultarse en el siguiente enlace: <https://www.itu-cop-guidelines.com/>

⁵³ Disponible en: <https://www.itu-cop-guidelines.com/implementation>

⁵⁴ Ritchie, H. “Australia prohíbe el uso de las redes sociales a menores de 16 años con la legislación de internet más estricta del mundo”, BBC, 24 de noviembre de 2024, disponible en: <https://www.bbc.com/mundo/articles/cq52v666vl3o>

⁵⁵ Hidalgo, D. *Retomar el control. 50 reflexiones para repensar nuestro futuro digital*, Libros de la Catarata, Madrid, 2024

3. La Asociación de la Adolescencia Libre de Móviles de la Comunidad Valencia como experiencia pionera de sensibilización y resistencia

La Asociación Adolescencia Libre de Móviles Comunidad Valenciana, surge el 20 de noviembre de 2023, como un movimiento asociativo ciudadano espontáneo, de la preocupación de un grupo de madres y padres, que a su vez son médicos, docentes, informáticos, juristas, académicos, etc. ante los perniciosos efectos que la digitalización masiva en el ámbito escolar y social genera en la salud física, mental y social.

Se trata de una asociación apolítica, aconfesional y plurilingüe, teniendo en cuenta la riqueza social del territorio donde se radica y que abarca a las tres provincias de Alicante, Castellón y Valencia. Su objetivo principal es a través del diálogo y la colaboración, concienciar a cuantas más familias, madres y padres y otros actores gubernamentales, educativos y sociales mejor, sobre los efectos negativos de las pantallas en menores desde su experiencia profesional y científica, por los riesgos que conlleva de manipulación o desinformación, entre otros, aunque también reconocen los beneficios de los avances tecnológicos empleados de una forma consciente, crítica y con cautela.

Ejemplo de esto, son sin duda sus orígenes que se remontan a la creación de un grupo de Telegram, donde se coordinan los participantes de las tres provincias y también con otros grupos similares a nivel nacional; a fin de compartir información y organizar sus acciones. En estos momentos más de 1.475 personas se han unido a este canal de Telegram y otras 675 al de Whatsapp. Una cifra que alcanza las 30.000 si tenemos en cuenta los canales de las diversas asociaciones similares presentes en otras autonomías

Posteriormente, a fin de dotarla de forma jurídica se constituye como asociación y se crea una página web, <https://aalmcv.es/>, en la que cualquier persona puede inscribirse de forma gratuita para ser miembro y conocer tanto su misión como sus acciones, cumplimentando un pequeño formulario y facilitando un correo electrónico a fin de poder recibir la información.

Entre los objetivos clave de la Asociación se pueden destacar los cinco siguientes:

- Concienciación temprana: así la asociación ayuda a divulgar y concienciar acerca de los efectos nocivos de cualquier tipo de pantalla, tanto a edad temprana como durante el desarrollo de los menores en su infancia y adolescencia, a través de la organización de charlas en colegios, vídeos e intercambio de material relevante en la materia como estudios científicos, médicos, novedades legislativas, etc.
- Apoyo a las familias: ofrece también apoyo para invertir la presión social y empoderar a las familias en cualquier contexto, tanto a nivel colegial, familiar e institucional, como por ejemplo la introducción de los denominados smartbooks en algunos centros escolares o la obligación de adquirir una tableta para poder realizar el seguimiento de las clases.

- Medidas en los centros escolares: por otra parte, también trata de mediar y consensuar medidas en los centros educativos que protejan de todo tipo de pantallas al menor devolviéndoles el bienestar y equilibrio que merecen para crecer, a tal fin ha mantenido reuniones informativas con Asociaciones de Padres en diferentes centros, Consejos Escolares y con la Conselleria de Educación y otros organismos como el Comité de Bioética de la Comunitat Valenciana
- Protección legislativa: finalmente desde su área jurídica tratan de plantear sus reivindicaciones ante las instituciones para que, mediante acciones legislativas, protejan a los menores de edad. Así algunos ejemplos de esta incidencia política e institucional fue la aprobación de la Resolución de 17 de abril de 2024, sobre determinados aspectos para la regulación del uso de dispositivos móviles en centros educativos no universitarios sostenidos con fondos públicos de la Comunitat Valenciana (DOGV 2024/3422, de 3 de mayo) o su participación en el proceso de consulta pública, el pasado mes de junio de 2025 sobre el Proyecto de orden de la Conselleria de Educación, Cultura, Universidades y Empleo, para la convivencia en los centros educativos no universitarios de la Comunitat Valenciana⁵⁶.
- Pacto de Familias, una herramienta disponible en la página web estatal de la entidad y de alcance nacional. Mediante este instrumento, cualquier padre, madre, tutor o representante legal de un menor puede adherirse indicando únicamente el centro educativo y el curso al que pertenece el alumno. Ya se han suscrito 10.628 pactos, que permiten a las familias visualizar de manera inmediata cuántos compañeros de la clase están comprometidos en retrasar la entrega del primer smartphone hasta los 16 años. Una iniciativa que busca reducir la presión social que padecen muchos padres y madres, fortaleciendo la protección de la infancia frente a una digitalización temprana y potencialmente nociva para su desarrollo.

Finalmente, entre las acciones y actividades desarrolladas en estos dos últimos años por la Asociación se pueden destacar las siguientes:

- Una de sus primeras acciones fue la publicación de un manifiesto fundacional de la Asociación en el que bajo el sugerente título “La vida es eso que pasa mientras estás mirando una pantalla” se exponen además de sus objetivos, los antecedentes que llevaron a su constitución, una serie de estrategias de limitación y actuación y finalmente unas conclusiones⁵⁷
- Pacto global de familias para retrasar la edad de entrega de un Smartphones a los menores hasta los 16 años⁵⁸ para el curso 2024/2025. Hasta la fecha el pacto ha sido suscrito por 10.488 personas y 2.654 centros.

⁵⁶ Unas aportaciones que pueden consultarse en el siguiente enlace:

<https://qvaparticipa.qva.es/legislation/processes/536/proposals/5998-limitacion-del-uso-de-pantallas>

⁵⁷ Este manifiesto puede consultarse también en: https://aalmcv.es/Manifiesto_AALMCV_ABRIL_2024.pdf

⁵⁸ El texto de este pacto y la adhesión al mismo se puede realizar en: <https://www.pacto.adolescencialibredemoviles.es/>

- Organización de charlas didácticas en colegios y otros espacios, como la organizada en Castalla impartida por Alejandro Fernández Vera, el 4 de marzo de 2024.
- Pic nics libres de móviles en espacios al aire libre para compartir experiencias y actividades alternativas de socialización a las pantallas.
- Campaña de adhesión del comercio para establecer rutas seguras y que permitan a los menores en caso de necesitarlo realizar una llamada.
- Participación en distintos eventos como la carrera de San Silvestre con una camiseta distintiva o en la Feria del Libro para dar visibilidad a la Asociación.
- Producción de una serie de videos⁵⁹ sobre cuestiones como: El Impacto de las Redes Sociales en la adolescencia, apaga el móvil, Villancico sin pantallas y otro material de sensibilización como corazones enamorados, campaña 8M, familias inteligentes, día internacional contra el maltrato infantil, el impacto de las redes sociales en la adolescencia, campaña de concienciación, pongamos en claro.
- Establecimiento de sinergias con otras entidades como Asafan, Familias Numerosas, Educación Digital Responsable, La niña amarilla para la prevención del suicidio y Family up. También está adherida a la propuesta de Pacto de Estado Menores digitales.

En último lugar cabe destacar que la Asociación no cuenta con ningún tipo de ayuda o subvención de carácter público o privado y que toda esta actividad se lleva a cabo gracias al compromiso y la solidaridad de sus miembros.

⁵⁹ Estos videos están disponibles en el siguiente canal de Youtube de la asociación:
<https://www.youtube.com/@AdolescenciaLibredeMovilesCV>

Bibliografía

Abadías, A. *Child Grooming: el embaucamiento de menores en la era del Metaverso y la Inteligencia Artificial*, Tecnos, Madrid, 2024.

Andrade, B., Guadix, I., Rial, A. y Suárez, F. Impacto de la tecnología en la adolescencia. Relaciones, riesgos y oportunidades, UNICEF España, Madrid, 2021.

Artigas, M. *Cerebro digital. Así nos transforma la tecnología*, National Geographic España, Madrid, 2025.

Bellver V. “Los derechos del niño en el entorno digital”, en Ballesteros, A. (ed.), *La digitocracia a debate*, Aranzadi, Pamplona, 2022, pp. 91-127.

Berrocal, A.I, “La protección jurídica de los menores y adolescentes ante el reto de la inteligencia artificial”, *La Ley Derecho de Familia: Revista jurídica sobre familia y menores*, Vol. 12, Nº. 46 (Monográfico: Inteligencia Artificial y familia: un cálido encuentro en plena encrucijada), 2025, pp. 4-32.

Brión, E. Protección de datos personales, redes sociales y menores de edad: Nuevos retos normativos en tiempos de *big data e inteligencia artificial*, Dykinson, Madrid, 2025.

Cóceres, M.C., “La inteligencia artificial como herramienta para la promoción de los derechos humanos de las personas mayores”, *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo*, Vol. 13, Nº. 2, 2025, pp. 103-124.

Costello, N., R. Sutton, M. Jones, M. Almassian, A. Raffoul, O. Ojumu, M. Salvia, M. Santoso, J. R. Kavanaugh, and S. B. Austin. 2023. “Algorithms, Addiction, and Adolescent Mental Health: An Interdisciplinary Study to Inform State-Level Policy Action to Protect Youth from the Dangers of Social Media.” *American Journal of Law & Medicine*, nº 49 (2-3), 2023, pp. 135–172

Febles, N. “Menores, derecho y algoritmo en la era de la inteligencia artificial”, *La Ley Derecho de Familia: Revista jurídica sobre familia y menores*, Vol. 12, Nº. 46 (Monográfico: Inteligencia Artificial y familia: un cálido encuentro en plena encrucijada), 2025, pp. 47-56.

Fernández Aller, C., Gutiérrez García, E. Puente, C. Águeda, Essinger, M. *Inteligencia artificial y derechos fundamentales. Menores, cuidado y vulnerabilidad digital*, Documentos de trabajo (Laboratorio de alternativas), Nº. 238, 2024

Goñi, E. “Inteligencia artificial y menores”, en *Digitalización del servicio público de justicia e inteligencia artificial judicial* / coord. por Donado, A., Jiménez Muñoz, F.J; Calaza, S. (dir.), Yáñez, F. (dir.), Dykinson, Madrid, 2024, pp. 153-176.

Hidalgo, D. *Anestesiados. La humanidad bajo el imperio de la tecnología*, Libros de la Catarata, Madrid, 2021

Hidalgo, D. *Retomar el control. 50 reflexiones para repensar nuestro futuro digital*, Libros de la Catarata, Madrid, 2024.

Pardo de Vera, M. “La incidencia de la Inteligencia Artificial en los derechos digitales de los menores. Regulación legal y problemática actual”, *La protección de la infancia en entornos digitales* / coord. por Rodríguez Rodríguez, D.M; España MartíM. (pr.), Dykinson, Madrid, 2024, pp. 123-130.

Rebollo, L. *Inteligencia artificial y derechos fundamentales*, Dykinson, Madrid, 2023.

Rosič, J., Schreurs, L., Janicke-Bowles, S., Vandenbosch, L. “Trajectories of digital flourishing in adolescence: The predictive roles of developmental changes and digital divide factors”, *Child Development*, Volume 95, Issue 5, 2024.

Santos, R. M. S., C. G. Mendes, G. Y. Sen Bressani, S. de Alcantara Ventura, Y. J. de Almeida Nogueira, D. M. de Miranda, and M. A. Romano-Silva, “The Associations between Screen Time and Mental Health in Adolescents: A Systematic Review.” *BMC Psychology* 11 (1), 2023, p. 127.

Sapien Labs, “The Youth Mind: Rising Aggression and Anger.” Rapid Report, 2025, disponible en: <https://doi.org/10.17605/OSF.IO/3WFJX>.

Thiagarajan,T., Newson, J. y Swaminathan, S. “Protecting the Developing Mind in a Digital Age: A Global Policy Imperative”, *Journal of Human Development and Capabilities*, nº 26:3, 2025, pp. 493-504.

Vanhoffelen, G., Gonzalez, A., Schreurs, L., Giraudeau, C., & Vandenbosch, L. (2025). “The Perfect Li(f)e: A Longitudinal Study on Positive Social Media Content and European Adolescents’ Perfectionism”. *Communication Research*, 0(0), 2025

Wu, D., Dong, X., Liu D.y Li H., “How Early Digital Experience Shapes Young Brains During 0-12 Years: A Scoping Review”, *Early Education and Development*, nº 7, 2023.

Itziar Artiñano. Neocolonialismo algorítmico y captura regulatoria: Desafíos jurídicos en la investigación farmacéutica global

Itziar Artiñano

Dotoranda UCM

[ACCEDE A LA GRABACIÓN](#)

Resumen

La irrupción de la inteligencia artificial (IA) en la investigación farmacéutica plantea desafíos que trascienden lo técnico para interpelar al Derecho internacional, la bioética y la justicia biomédica. Su uso en ensayos clínicos globales configura lo que aquí se denomina neocolonialismo algorítmico: un patrón de dominación digital mediante el cual datos y sujetos del Sur global se convierten en insumo para corporaciones del Norte. Unido a la captura regulatoria, este fenómeno erosiona la soberanía de los Estados y compromete la legitimidad de la industria farmacéutica. Mediante un análisis jurídico-comparado y crítico, que integra dimensiones normativas y éticas, se estudian casos paradigmáticos —como el ensayo de Pfizer en Nigeria (1996)— junto con los principales marcos regulatorios vigentes: Reglamento (UE) 536/2014 sobre ensayos clínicos, Reglamento General de Protección de Datos (2016/679), Ley de Inteligencia Artificial (UE) 2024/1689 y directrices indias de 2023 sobre IA sanitaria. Se identifican dos fracturas esenciales: la insuficiencia extraterritorial de los estándares europeos y la vulnerabilidad estructural de países con instituciones débiles frente a imperativos tecnológicos y corporativos, en un contexto carente de tratado internacional vinculante. La aportación consiste en delinear un modelo normativo integral de gobernanza algorítmica transnacional, sustentado en los principios de precaución, proporcionalidad, soberanía sanitaria y justicia epistémica. Entre las medidas propuestas figuran registros públicos de algoritmos, auditorías independientes, fortalecimiento de agencias regulatorias en el Sur global y mecanismos internacionales de reparación. La inteligencia artificial aplicada a los ensayos clínicos globales representa hoy una encrucijada histórica: reproducir las lógicas de dominación o inaugurar un paradigma de equidad. La consolidación de un marco jurídico internacional robusto constituye, por ello, condición de posibilidad para erigir la justicia sanitaria internacional como principio rector de la biomedicina en la era digital. Se sostiene que la gobernanza algorítmica global debe transformarse en un marco jurídico vinculante basado en la justicia sanitaria internacional.

Palabras clave: captura regulatoria, Derecho internacional, ensayos clínicos, gobernanza sanitaria, inteligencia artificial, justicia sanitaria internacional, neocolonialismo algorítmico.

Abstract

The emergence of artificial intelligence (AI) in pharmaceutical research raises challenges that transcend the technical sphere to question international law, bioethics, and biomedical justice. Its deployment in global clinical trials configures what is here termed algorithmic neocolonialism: a pattern of digital domination through which data and subjects from the Global South become mere inputs for corporations in the Global North. Coupled with regulatory capture, this phenomenon erodes state sovereignty and undermines the legitimacy of the pharmaceutical industry. Through a comparative and critical legal analysis integrating normative and ethical dimensions, paradigm cases—such as the Pfizer trial in Nigeria (1996)—are examined alongside the main regulatory frameworks currently in force: EU Regulation 536/2014 on clinical trials, the General Data Protection Regulation (2016/679), the Artificial Intelligence Act (EU) 2024/1689, and India’s 2023 guidelines on health-related AI. Two structural fractures are identified: the extraterritorial insufficiency of European standards and the institutional vulnerability of states with fragile systems in the face of technological and corporate imperatives, within a context devoid of any binding international treaty. The contribution lies in outlining an integral normative model of transnational algorithmic governance, grounded in the principles of precaution, proportionality, health sovereignty, and epistemic justice. Proposed measures include public registries of algorithms, independent audits, the strengthening of regulatory agencies in the Global South, and international mechanisms of redress. Artificial intelligence applied to today global clinical trials represents a historical crossroads: to reproduce entrenched logics of domination or to inaugurate a paradigm of equity. The consolidation of a robust international legal framework thus constitutes a necessary condition for establishing international health justice as the guiding principle of biomedicine in the digital age. Global algorithmic governance must evolve into a binding legal order grounded in international health justice.

Keywords: algorithmic neocolonialism, artificial intelligence, clinical trials, health governance, international health justice, international law, regulatory capture.

1. Introducción

Las transformaciones de la investigación farmacéutica han estado históricamente ligadas a revoluciones tecnológicas: de la química sintética a la biología molecular, de la genómica a la biomedicina personalizada. La irrupción de la inteligencia artificial (IA) inaugura, sin embargo, un cambio de naturaleza distinta. No se trata ya de una herramienta auxiliar de descubrimiento, sino de un sistema de decisión autónoma capaz de reconfigurar integralmente las fases de investigación y desarrollo: desde la identificación de moléculas hasta la selección de sujetos y la interpretación de resultados en los ensayos clínicos. Esta capacidad otorga a la IA un poder inédito en la economía política de la salud, con promesas de eficiencia sin precedentes, pero también con riesgos de dominación y exclusión de magnitud igualmente inédita.

El problema central no reside únicamente en la opacidad de algoritmos propietarios ni en los sesgos derivados de datos entrenados en poblaciones no representativas, sino en la manera en que estas tecnologías se insertan en una geografía desigual de la investigación biomédica. La externalización de ensayos hacia jurisdicciones con estándares más débiles ya había generado acusaciones de colonialismo sanitario; la IA añade ahora una capa de sofisticación que puede ocultar y amplificar esas prácticas. Es lo que cabe denominar neocolonialismo algorítmico: un régimen en el que datos y cuerpos del Sur global se convierten en materia prima para sistemas diseñados y controlados en el Norte, bajo la legitimación aparente de la objetividad técnica.

A esta dinámica se suma la captura regulatoria, un fenómeno descrito desde la teoría económica del derecho, pero que en el ámbito farmacéutico adquiere dimensiones singulares: agencias debilitadas o cooptadas, incapaces de fiscalizar tecnologías cuya complejidad técnica las hace depender de la propia industria regulada. El resultado es un doble déficit de legitimidad: por un lado, Estados que no logran proteger a sus ciudadanos; por otro, corporaciones que operan sin un régimen claro de responsabilidad por los efectos de sus algoritmos.

La paradoja es clara: mientras la Unión Europea promulga normas pioneras —el Reglamento 536/2014 sobre ensayos clínicos, el RGPD (2016/679) y la reciente Ley de Inteligencia Artificial (2024/1689)—, su eficacia se ve limitada por la falta de extraterritorialidad y por lagunas específicas en el ámbito biomédico. En ausencia de un tratado internacional vinculante, los vacíos se multiplican en aquellos contextos donde la protección debería ser mayor.

En este escenario, la cuestión jurídica ya no es solo cómo regular la IA como tecnología o los ensayos como práctica biomédica, sino cómo articular un marco integral que responda a la intersección de ambos campos. La respuesta exige concebir una gobernanza algorítmica transnacional, cimentada en principios de precaución, proporcionalidad, soberanía sanitaria y justicia epistémica, capaz de resistir la instrumentalización de la IA como dispositivo de dominación y de reorientarla hacia la realización de una auténtica justicia sanitaria

internacional. Desde un enfoque jurídico-comparado y crítico, se examina la interacción entre inteligencia artificial y ensayos clínicos globales, con el objetivo de delinear un modelo normativo transnacional de gobernanza algorítmica orientado a la protección de los derechos humanos y a la equidad sanitaria internacional.

2. IA en la investigación farmacéutica: automatización, opacidad y sesgos

La irrupción de la inteligencia artificial en la investigación farmacéutica representa un punto de inflexión que excede la mera modernización técnica: constituye una reconfiguración del poder decisional en el ciclo de desarrollo de fármacos y, en particular, de los ensayos clínicos internacionales. Allí donde antes predominaba la deliberación científica y ética, emerge ahora un entramado algorítmico que orienta la identificación de moléculas, la selección de sujetos, la monitorización de datos y la evaluación de resultados. Estas funciones, presentadas bajo la promesa de eficiencia, coste reducido y precisión diagnóstica, plantean una cuestión decisiva: si la ciencia biomédica será gobernada por criterios de transparencia y justicia o, por el contrario, por lógicas de opacidad tecnológica y dominación corporativa.

El primer eje problemático se sitúa en la opacidad algorítmica. La industria farmacéutica ampara bajo el secreto comercial sistemas de decisión que, sin embargo, afectan derechos fundamentales de los participantes. La exclusión de un paciente de un ensayo, la evaluación de su riesgo o la identificación de eventos adversos pueden depender de procesos no auditables por investigadores independientes ni por autoridades regulatorias⁶⁰. En este contexto, la tan invocada *explicabilidad algorítmica* se convierte en un requisito ético-jurídico ineludible: sin ella, se vacía de contenido el consentimiento informado y se erosiona el principio de igualdad de trato⁶¹. La reciente Ley de Inteligencia Artificial de la Unión Europea (2024/1689) reconoce el carácter de “alto riesgo” de estas aplicaciones⁶², pero lo hace con disposiciones vagas y sin garantizar una validación independiente efectiva, lo que abre un nuevo terreno de incertidumbre regulatoria.

⁶⁰ 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 (RGPD). *Diario Oficial de la Unión Europea*, L 119, 4 mayo 2016, p. 1-88.

⁶¹ ASOCIACIÓN MÉDICA MUNDIAL. *Declaración de Helsinki. Principios éticos para las investigaciones médicas en seres humanos*. Fortaleza, 2013. Disponible en: <https://www.wma.net/es/polices-post/declaracion-de-helsinki-de-la-amm-principios-eticos-para-las-investigaciones-medicas-en-seres-humanos/>

⁶² REGLAMENTO (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 12 de julio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial). *Diario Oficial de la Unión Europea*, L 237, 12 julio 2024, p. 1-142.

El segundo eje lo constituyen los sesgos algorítmicos, cuya relevancia se proyecta con particular fuerza en clave Norte-Sur. Los modelos entrenados mayoritariamente con datos clínicos de poblaciones europeas o norteamericanas generan predicciones deficientes en contextos del Sur global, donde factores genéticos, ambientales y sociales difieren de manera sustancial. La OMS ha advertido que ello puede traducirse en exclusiones injustificadas, diagnósticos erróneos o monitorizaciones insuficientes⁶³, replicando —ahora bajo ropaje digital— la histórica infrarepresentación de mujeres, minorías étnicas y personas mayores en la investigación clínica. Esta dinámica no es un mero fallo técnico: constituye una forma de discriminación algorítmica estructural que perpetúa la desigualdad biomédica y agrava lo que la teoría crítica ha denominado injusticia epistémica⁶⁴.

El tercer eje problemático es la dilución de la responsabilidad humana. La delegación de decisiones críticas en sistemas automatizados —como la continuidad o interrupción de un ensayo a partir de resultados interinos— plantea interrogantes esenciales sobre la imputación jurídica de los daños. La confianza ciega en “lo que dicta el algoritmo” erosiona principios éticos consolidados, como el deber del investigador de priorizar en todo momento la seguridad de los sujetos. Algunos ordenamientos, como el de India, han reaccionado estableciendo que la responsabilidad legal debe definirse antes de la implementación⁶⁵, con posibilidad de imputar directamente a los diseñadores de IA los daños ocasionados por fallos tecnológicos. Este modelo apunta hacia una noción de responsabilidad objetiva por riesgo tecnológico, imprescindible para evitar zonas de impunidad transnacional⁶⁶. En el plano jurídico, esta delegación reabre el debate sobre la imputación de daños en entornos algorítmicos. La Propuesta de Directiva relativa a la responsabilidad por inteligencia artificial (COM(2022) 496 final) avanza hacia un modelo de responsabilidad por riesgo tecnológico, que podría adaptarse al contexto biomédico mediante la obligación de demostrar diligencia en la validación, supervisión y trazabilidad de los algoritmos empleados en los ensayos clínicos. De este modo, se preserva el principio clásico de imputabilidad sin desincentivar la innovación.

En suma, la IA aplicada a la investigación farmacéutica inaugura un régimen ambivalente: por un lado, amplía las posibilidades de innovación biomédica; por otro, multiplica los riesgos de

⁶³ ORGANIZACIÓN MUNDIAL DE LA SALUD. *Ethics and governance of artificial intelligence for health*. Ginebra: OMS, 2021. Disponible en: <https://www.who.int/publications/i/item/9789240029200>

⁶⁴ BENJAMIN, Ruha; MULDOON, Julianna y otros. «Artificial Intelligence in the Colonial Matrix of Power». *Philosophy & Technology*, 2023, vol. 36, n.º 87. Disponible en <https://link.springer.com/article/10.1007/s13347-023-00687-8#citeas> :

⁶⁵ INDIAN COUNCIL OF MEDICAL RESEARCH (ICMR). *Ethical guidelines for application of artificial intelligence in biomedical research and healthcare*. Nueva Delhi: ICMR, 2023. Disponible en <https://www.icmr.gov.in/>

⁶⁶ COMISIÓN EUROPEA. *Propuesta de Directiva relativa a la responsabilidad por inteligencia artificial*. COM(2022) 496 final, Bruselas, 28 septiembre 2022.

opacidad, sesgo y desresponsabilización. En contextos globalmente desiguales, estas dinámicas no son neutras: se insertan en un patrón de neocolonialismo algorítmico, mediante el cual la automatización puede convertirse en instrumento de extracción de valor y de reproducción de asimetrías estructurales. La cuestión central, desde la óptica del Derecho internacional y de la justicia sanitaria, no es ya si la IA puede mejorar la investigación clínica, sino cómo garantizar que no consolide bajo nuevas formas las lógicas de dominación que han acompañado históricamente a la biomedicina global.

3. Ensayos clínicos globales y neocolonialismo sanitario digital

La globalización de la investigación farmacéutica ya arrastraba profundas controversias éticas antes de la irrupción de la inteligencia artificial. Desde la década de 1990, las grandes farmacéuticas comenzaron a externalizar ensayos clínicos hacia países de economías emergentes o en desarrollo, atraídas por menores costes, escasa densidad normativa y la disponibilidad de pacientes “tratamiento-ingenuos”. La antropóloga Adriana Petryna describió este fenómeno bajo la noción de *ethical variability*, es decir, la instrumentalización de la desigualdad regulatoria global como ventaja competitiva⁶⁷. Se trata, en términos jurídicos, de un forum shopping ético, donde las corporaciones eligen la jurisdicción más permisiva para realizar prácticas que no serían aceptables en sus países de origen.

Un caso paradigmático fue el ensayo de Pfizer con trovafloxacino (Trovan) en Kano (Nigeria, 1996), en plena epidemia de meningitis. La compañía administró el fármaco experimental a unos 200 menores sin consentimiento informado válido ni autorización ética robusta⁶⁸. Posteriores investigaciones revelaron la falsificación de documentos de aprobación y la continuidad del ensayo incluso en pacientes que empeoraban, pese a existir tratamiento estándar disponible. Once niños fallecieron y decenas sufrieron secuelas irreversibles. El gobierno nigeriano declaró ilegal el ensayo, y el litigio internacional culminó con un acuerdo extrajudicial de 75 millones de dólares bajo la Alien Tort Statute en Estados Unidos, sin admisión de responsabilidad⁶⁹. Más allá de los daños directos, el episodio erosionó la confianza

⁶⁷ PETRYNA, Adriana. *When Experiments Travel: Clinical Trials and the Global Search for Human Subjects*. Princeton: Princeton University Press, 2009, vol. 5, n° 2. Disponible en: <https://scispace.com/pdf/when-experiments-travel-clinical-trials-and-the-global-1pkd95gsut.pdf>

⁶⁸ WISE, Jacqueline. *Pfizer accused of testing new drug without ethical approval*. BMJ (Clinical Research ed.), 2001, vol. 322, n° 7280, p. 194. ISSN 0959-8138. DOI: <https://doi.org/10.1136/bmj.322.7280.194>

⁶⁹ ESTADOS UNIDOS. Corte de Apelaciones del Segundo Circuito. *Abdullahi v. Pfizer, Inc.*, 562 F.3d 163 (2d Cir. 2009), *cert. denied*, 130 S. Ct. 3541 (2010).

de la población en la medicina occidental, alimentando narrativas de colonialismo médico y boicots a campañas de vacunación posteriores.

Este caso revela dos fracturas estructurales: por un lado, la distancia entre principios éticos universales —como los recogidos en la Declaración de Helsinki (2013)⁷⁰ y las Pautas CIOMS (2016)⁷¹— y su aplicación efectiva en contextos de débil institucionalidad; por otro, la insuficiencia de los mecanismos internacionales de responsabilidad para tutelar los derechos de poblaciones vulnerables.

La irrupción de la inteligencia artificial añade ahora una dimensión inédita. Los sistemas algorítmicos permiten identificar poblaciones “óptimas” a través de cribados digitales y coordinar ensayos a distancia, sin necesidad de mediación humana visible. Este proceso facilita el control remoto de investigaciones en países con regulaciones laxas y, al mismo tiempo, encubre decisiones sensibles bajo la supuesta objetividad de la tecnología. La consecuencia es una reconfiguración del poder colonial: lo que antes requería presencia física y negociación local puede hoy ejecutarse mediante plataformas digitales que operan desde centros corporativos del Norte.

Se trata de lo que la literatura crítica ha denominado colonialismo de datos (*data colonialism*) y, en su dimensión biomédica, biocolonialismo⁷²: la extracción masiva de datos y muestras biológicas de poblaciones del Sur global para alimentar algoritmos cuyos beneficios — fármacos, patentes, publicaciones— se concentran en los mercados del Norte. Esta transferencia unidireccional de valor se ve agravada por la opacidad algorítmica, que dificulta el escrutinio de decisiones de inclusión, exclusión o seguimiento de pacientes en ensayos clínicos.

El impacto no es meramente técnico. Este neocolonialismo sanitario digital compromete derechos reconocidos en el Pacto Internacional de Derechos Económicos, Sociales y Culturales (PIDESC) —el derecho a la salud (art. 12), el derecho a beneficiarse del progreso científico (art. 15) y el principio de igualdad y no discriminación (art. 2)⁷³—. Al relegar a comunidades del Sur

⁷⁰ WORLD MEDICAL ASSOCIATION. *Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*. Fortaleza: 64th WMA General Assembly, 2013.

⁷¹ COUNCIL FOR INTERNATIONAL ORGANIZATIONS OF MEDICAL SCIENCES (CIOMS). *International Ethical Guidelines for Health-Related Research Involving Humans*. Ginebra: CIOMS/OMS, 2016. ISBN 9789290360889.

⁷² RAJAN, Kaushik Sunder. *Biocapital: The Constitution of Postgenomic Life*. Durham: Duke University Press, 2006. ISBN 9780822337349

⁷³ NACIONES UNIDAS. *Pacto Internacional de Derechos Económicos, Sociales y Culturales*. Resolución 2200 A (XXI), adoptada el 16 de diciembre de 1966; entrada en vigor, 3 de enero de 1976. *United Nations Treaty Series*, vol. 993, p. 3. Disponible en: <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights>

al rol de meros proveedores de datos y cuerpos, mientras el Norte monopoliza la interpretación y apropiación de conocimiento, se perpetúa una injusticia epistémica que consolida asimetrías históricas bajo la apariencia de innovación neutral⁷⁴.

En suma, la inteligencia artificial no solo intensifica los patrones coloniales de la investigación biomédica global, sino que los automatiza y legitima mediante dispositivos técnicos difíciles de auditar. Sin garantías de transparencia, participación y soberanía sanitaria, corremos el riesgo de consolidar un modelo de colonialismo sanitario digital, en el que la vulnerabilidad estructural de las poblaciones se convierte en la materia prima del progreso farmacéutico global.

4. Captura regulatoria en entornos de baja regulación

La captura regulatoria describe el fenómeno por el cual las agencias llamadas a supervisar una industria terminan actuando en beneficio de ésta y no del interés público. En el ámbito farmacéutico global, esta dinámica adopta modalidades múltiples: desde la redacción de marcos normativos influenciados por el lobby corporativo —incluidos tratados comerciales que refuerzan la exclusividad de los datos de ensayo clínico frente a medicamentos genéricos⁷⁵— hasta la inacción de reguladores condicionados por la dependencia económica de la inversión extranjera.

El caso de Nigeria en 1996, con el ensayo del trovafloxacin de Pfizer, ilustra crudamente este vacío: en ausencia de una ley que exigiera aprobación previa por un comité ético independiente, la compañía obtuvo autorizaciones administrativas cuestionables y operó sin una supervisión real⁷⁶. De manera análoga, los escándalos suscitados en India a inicios de los 2000 evidenciaron la inscripción masiva de pacientes vulnerables en ensayos sin su consentimiento informado, lo que llevó a la Corte Suprema en 2013 a exigir reformas profundas de las Drugs and Cosmetics Rules⁷⁷. Estos episodios reflejan un patrón: en el Sur global, los avances normativos en investigación clínica surgen casi siempre como reacción a abusos consumados.

⁷⁴ ESTADOS UNIDOS. *Alien Tort Statute*, 28 U.S.C. § 1350. Washington, D.C.: Office of the Law Revision Counsel, [ed. vigente]. Disponible en: <https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title28-section1350>

⁷⁵ CORREA, Carlos María. *Intellectual Property and Public Health: The TRIPS Agreement and Patents on Medicines*. Ginebra: International Centre for Trade and Sustainable Development (ICTSD), 2000. Disponible en: https://www.researchgate.net/publication/11923311_Health_and_intellectual_property_rights

⁷⁶ ABDULLAHI v. Pfizer, Inc. 562 F.3d 163 (2d Cir. 2009). United States Court of Appeals for the Second Circuit. Disponible en: <https://casetext.com/case/abdullahi-v-pfizer-inc-2>

⁷⁷ SUPREME COURT OF INDIA. *Swasthya Adhikar Manch and Others v. Union of India & Others*, Writ Petition (Civil) No. 33 of 2012, sentencia del 3 de enero de 2013. Disponible en: <https://indiankanoon.org/doc/11325454/>

La irrupción de la inteligencia artificial reconfigura estos mecanismos de captura. A la clásica subordinación económica y política se añade la captura cognitiva: la brecha técnica entre reguladores y regulados impide a muchas agencias auditar algoritmos de reclutamiento o análisis de datos, quedando a merced de la información proporcionada por la propia empresa⁷⁸. La narrativa del “progreso tecnológico” refuerza esta asimetría, pues gobiernos deseosos de proyectar modernidad pueden mostrarse proclives a flexibilizar requisitos, aceptando estudios con IA sin la debida fiscalización.

El fenómeno también opera a escala transnacional. La inclusión de cláusulas de data exclusivity en tratados comerciales ilustra cómo intereses corporativos logran blindajes normativos que limitan la capacidad de escrutinio en los países receptores⁷⁹. Nada impide que dinámicas análogas emerjan en la regulación internacional de la IA aplicada a la salud: si los estándares globales son capturados por los actores dominantes, se consolidará un marco voluntarista y laxo, en detrimento de obligaciones vinculantes centradas en los derechos humanos.

Un exponente particularmente problemático es el de las Contract Research Organizations (CROs), que se publicitan como expertas en “navegar” regulaciones locales, explotando las brechas normativas. La proliferación de estas entidades tras procesos de desregulación produjo una competencia a la baja en garantías éticas: estudios empíricos han mostrado que en numerosos ensayos realizados en países en desarrollo con participación de CROs, menos de la mitad fue revisada por un comité ético independiente⁸⁰.

Desde la perspectiva del Derecho internacional, la captura regulatoria vulnera directamente el artículo 12 del Pacto Internacional de Derechos Económicos, Sociales y Culturales y la Observación General nº 14 del Comité DESC, que imponen a los Estados la obligación de proteger a los individuos frente a injerencias de actores privados que puedan comprometer su derecho a la salud. Asimismo, contradice los principios de la Declaración de Helsinki y de las Pautas CIOMS, que exigen un control ético independiente y el respeto irrestricto de la dignidad humana⁸¹.

⁷⁸ WORLD HEALTH ORGANIZATION. *Ethics and governance of artificial intelligence for health: WHO guidance*. Ginebra: World Health Organization, 2021. ISBN 978-92-4-002920-0. Disponible en: <https://www.who.int/publications/i/item/9789240029200>

⁷⁹ ROVIRA, Joan; KILIC, Burcu. *Data exclusivity and access to medicines*. Ginebra: South Centre, 2019. Disponible en: <https://www.southcentre.int/research-paper-101-october-2019/>

⁸⁰ GOLDRACRE, Ben. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. London: Fourth Estate, 2012. ISBN 978-0-00-735074-2.

⁸¹ COUNCIL FOR INTERNATIONAL ORGANIZATIONS OF MEDICAL SCIENCES (CIOMS); WORLD HEALTH ORGANIZATION (WHO). *International Ethical Guidelines for Health-related Research Involving Humans*. Ginebra: CIOMS; 2016. ISBN 978-92-

En definitiva, la captura regulatoria en entornos de baja regulación no constituye solo una deficiencia administrativa: implica la cesión de soberanía sanitaria y la normalización de un desequilibrio estructural en el reparto de riesgos biomédicos. Allí donde los Estados consienten esta captura sin resistencia, no solo se degrada la protección de los sujetos de investigación, sino que se socava la legitimidad misma del orden jurídico internacional, que proclama la primacía de la dignidad humana sobre los intereses económicos.

5. Límites de la regulación vigente: Unión Europea y estándares internacionales

Las normas internacionales de referencia en investigación clínica —en particular las Guías de Buenas Prácticas Clínicas del ICH-GCP (E6 [R2] y borrador E6 [R3])⁸², las Consideraciones Generales del ICH E8 (R1)⁸³ y la norma ISO 14155:2020⁸⁴, junto con los estándares de integridad de datos GxP/ALCOA+⁸⁵— se presentan habitualmente como garantía de cumplimiento ético y científico. Sin embargo, en el contexto algorítmico global muestran limitaciones sustanciales.

El enfoque de gestión del riesgo promovido por ICH-GCP se apoya cada vez más en sistemas de inteligencia artificial encargados de seleccionar sujetos, detectar señales o priorizar auditorías⁸⁶. Cuando estos algoritmos son opacos o inaccesibles para los reguladores, la llamada “supervisión humana” se convierte en una formalidad, y resulta imposible verificar la validez de las decisiones automatizadas. De manera análoga, ICH E8 (R1) exige definir “lo que importa para la calidad”, lo que en entornos de IA implicaría especificar la finalidad del sistema, los datos de entrenamiento y los sesgos identificados. En la práctica, pocos marcos normativos exigen esa información o contemplan evaluaciones independientes, lo que reduce la observancia a un ejercicio meramente declarativo.

9036-088-9. Disponible en: <https://cioms.ch/publications/product/international-ethical-guidelines-for-health-related-research-involving-humans/>

⁸² INTERNATIONAL COUNCIL FOR HARMONISATION (ICH). E6(R2) Guideline for Good Clinical Practice; Draft E6(R3). Ginebra: ICH, 2016 y 2023. Disponible en <https://www.ich.org/page/efficacy-guidelines>

⁸³ INTERNATIONAL COUNCIL FOR HARMONISATION (ICH). E8(R1) General Considerations for Clinical Studies. Ginebra: ICH, 2019. Disponible en <https://www.ich.org/page/efficacy-guidelines>

⁸⁴ INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). ISO 14155:2020 Clinical investigation of medical devices for human subjects — Good clinical practice. Ginebra: ISO, 2020. Disponible en: <https://www.iso.org/standard/71690.html>

⁸⁵ MEDICINES AND HEALTHCARE PRODUCTS REGULATORY AGENCY (MHRA). GxP Data Integrity Guidance and Definitions. Londres: MHRA, marzo 2018 (actualizado 2021). Disponible en: <https://www.gov.uk/government/publications/gxp-data-integrity-guidance-and-definitions>

⁸⁶ **EUROPEAN MEDICINES AGENCY**. Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. Amsterdam: EMA, 9 septiembre 2024. Documento EMA/CHMP/CVMP/83833/2023. Disponible en: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf

Los principios GxP y las normas ISO establecen obligaciones de trazabilidad y conservación íntegra de los datos, pero su aplicación resulta problemática cuando los procesos se externalizan a CROs o proveedores tecnológicos que operan en nubes extraterritoriales. Aunque el patrocinador sigue siendo responsable del ensayo conforme a ICH-GCP, esa responsabilidad se diluye si no existen cláusulas contractuales que aseguren derechos de auditoría, acceso a la documentación técnica del modelo y control sobre la localización de los datos. Esta dependencia tecnológica genera lo que puede denominarse captura cognitiva, al situar a las agencias de supervisión en una posición subordinada respecto de los desarrolladores de IA.

En suma, los marcos ICH e ISO siguen siendo necesarios, pero resultan insuficientes para garantizar una investigación clínica ética en la era de la inteligencia artificial. Su actualización debería incorporar exigencias de explicabilidad algorítmica, evaluaciones externas de validez y sesgo, mecanismos de reparación y soberanía de datos, de modo que el cumplimiento formal no oculte prácticas de externalización que perpetúan las asimetrías estructurales entre Norte y Sur. En la actualidad, no existe un tratado internacional vinculante que regule de manera específica la ética de la inteligencia artificial en la investigación biomédica. La gobernanza global se articula en torno a un mosaico fragmentado de normas nacionales y regionales, complementado por instrumentos de *soft law* de alcance universal, cuyo cumplimiento depende en última instancia de la voluntad de los Estados.

En el contexto europeo, la protección de los sujetos de investigación se ha reforzado de forma notable. El Reglamento (UE) 536/2014 sobre ensayos clínicos, plenamente aplicable desde 2022, impuso mayores exigencias de transparencia y seguridad. Entre sus aportaciones destaca la base de datos pública CTIS, que obliga a registrar y reportar ensayos autorizados, con el fin de evitar prácticas históricas de ocultación de resultados negativos, denunciadas por autores como Goldacre⁸⁷. Asimismo, se consolidó la uniformidad en el consentimiento informado, se fortaleció el papel de los comités de ética y se establecieron reglas estrictas para la notificación de eventos adversos. Sin embargo, este marco carece de extraterritorialidad: las farmacéuticas europeas que realizan ensayos fuera de la UE no están obligadas a replicar estos estándares, salvo que pretendan utilizar los datos en solicitudes de autorización de comercialización ante la Agencia Europea de Medicamentos. Ello genera un doble estándar que permite trasladar fases de investigación a jurisdicciones más permisivas, en abierta contradicción con el espíritu

⁸⁷ GOLDRACRE, Ben. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. London: Fourth Estate, 2012. ISBN 978-0-00-735074-2.

universal de la Declaración de Helsinki, que exige priorizar siempre el bienestar de los sujetos sobre los intereses de la ciencia y la sociedad⁸⁸.

A este marco se suma el Reglamento General de Protección de Datos (RGPD, 2016/679), que otorga amplias garantías sobre el tratamiento de datos personales sensibles, incluidos los de salud. Principios como la limitación de la finalidad, la minimización o la responsabilidad proactiva constituyen un avance significativo, y en teoría las farmacéuticas europeas deben cumplirlos incluso en ensayos realizados fuera de la UE⁸⁹. No obstante, la existencia de excepciones para la “investigación científica” abre espacios de ambigüedad, y el RGPD no aborda de forma directa la opacidad algorítmica: el artículo 22 sobre decisiones automatizadas tiene un alcance limitado y difícilmente cubre procesos como la selección de cohortes mediante IA, que suelen presentarse como meramente auxiliares al juicio humano⁹⁰.

El reciente Reglamento (UE) 2024/1689, Ley de Inteligencia Artificial, constituye un hito al establecer un marco integral para el uso de IA en el mercado europeo. Su sistema de clasificación por niveles de riesgo sitúa a las aplicaciones sanitarias críticas en la categoría de “alto riesgo”, lo que conlleva requisitos obligatorios de calidad, gestión de riesgos, transparencia y supervisión humana⁹¹. Este enfoque supone un avance en la tutela de derechos fundamentales frente a los riesgos tecnológicos, pero también revela limitaciones: su aplicación se restringe a la UE y su ámbito de influencia internacional es débil, mientras que persisten lagunas respecto de la exigencia de explicabilidad en algoritmos complejos y de la interacción con otros marcos como el Reglamento de Productos Sanitarios⁹². En consecuencia, aunque constituye un referente normativo, su capacidad de incidir en prácticas extraterritoriales sigue siendo reducida.

Fuera de Europa, la regulación descansa en marcos éticos no vinculantes. La Declaración de Helsinki de 2013 y las Pautas CIOMS de 2016 continúan siendo referencias cardinales, al establecer principios de justicia, transparencia y protección reforzada de poblaciones

⁸⁸ WORLD MEDICAL ASSOCIATION (WMA). *Declaración de Helsinki: Principios éticos para las investigaciones médicas en seres humanos*. Fortaleza: 64ª Asamblea General, octubre 2013. Disponible en: <https://www.wma.net/es/policies-post/declaracion-de-helsinki-de-la-amm-principios-eticos-para-las-investigaciones-medicas-en-seres-humanos/>

⁸⁹ EUROPEAN UNION. *Reglamento (UE) 2016/679, General Data Protection Regulation (GDPR)*. Diario Oficial de la Unión Europea, L 119, 4 mayo 2016. Disponible en: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

⁹⁰ KAMINSKI, Margot E.; MALGIERI, Gianclaudio. *Algorithmic impact assessments under the GDPR: producing multi-layered explanations*. *International Data Privacy Law*, 2021, vol. 11, nº 2, pp. 125-144. DOI: 10.1093/idpl/ipaa020.

⁹¹ EUROPEAN UNION. *Reglamento (UE) 2024/1689, Artificial Intelligence Act*. Diario Oficial de la Unión Europea, L, 12 julio 2024. Disponible en: <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32024R1689>

⁹² VEALE, Michael; ZUIDERVEEN BORGESIU, Frederik. *Demystifying the Draft EU Artificial Intelligence Act*. *Computer Law Review International*, 2021, vol. 22, nº 4, pp. 97-112. DOI: 10.9785/cr-2021-220402.

vulnerables⁹³. Sin embargo, su eficacia depende de la voluntad de los Estados de incorporarlas en sus legislaciones y de la solidez de los comités de ética locales. La OMS, en su informe de 2021 sobre ética y gobernanza de la IA en salud, añadió principios esenciales como la autonomía, la rendición de cuentas y la inclusividad, mientras que la UNESCO, en ese mismo año, aprobó su *Recomendación sobre la ética de la IA*, aplicable a todo tipo de usos tecnológicos⁹⁴. Ambos instrumentos enriquecen el marco conceptual, pero carecen de fuerza jurídica vinculante.

Algunos Estados han comenzado a reaccionar mediante iniciativas nacionales. India constituye un ejemplo notable con sus Directrices éticas de 2023 para la IA en sanidad, que exigen transparencia algorítmica, consentimiento digital específico y auditorías para prevenir sesgos, además de imponer el registro obligatorio de ensayos clínicos y, desde 2025, de las organizaciones de investigación por contrato⁹⁵. Estas medidas apuntan a consolidar la soberanía regulatoria y evitar que el país se convierta en un espacio de experimentación desprotegido.

No obstante, la mayoría de países con gran volumen de ensayos clínicos —particularmente en África, Sudeste Asiático y América Latina— carecen de leyes específicas sobre IA y, en muchos casos, de marcos robustos sobre ensayos clínicos. Esta situación perpetúa un vacío normativo que permite a las farmacéuticas desplegar un doble estándar: un estricto cumplimiento en Europa, junto con prácticas más laxas en territorios con menor capacidad de control. En definitiva, el régimen actual reproduce desigualdades estructurales y revela la urgencia de un tratado internacional vinculante que establezca garantías universales frente a los riesgos de la inteligencia artificial en la investigación biomédica.

6. Hacia una gobernanza algorítmica transnacional

Frente a los desafíos descritos, resulta indispensable diseñar una **gobernanza jurídica transnacional** de la investigación farmacéutica en la era de la inteligencia artificial. La naturaleza global de los ensayos clínicos y la complejidad de las tecnologías algorítmicas hacen insuficiente

⁹³ COUNCIL FOR INTERNATIONAL ORGANIZATIONS OF MEDICAL SCIENCES (CIOMS). *International Ethical Guidelines for Health-related Research Involving Humans*. Ginebra: CIOMS, 2016. Disponible en: <https://cioms.ch/publications/product/international-ethical-guidelines-for-health-related-research-involving-humans/>

⁹⁴ WORLD HEALTH ORGANIZATION. *Ethics and governance of artificial intelligence for health: WHO guidance*. Ginebra: World Health Organization, 2021. ISBN 978-92-4-002920-0. Disponible en: <https://www.who.int/publications/i/item/9789240029200>

⁹⁵ INDIAN COUNCIL OF MEDICAL RESEARCH (ICMR). *Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare*. New Delhi: ICMR, 2023. Disponible en: https://www.icmr.gov.in/icmrobject/uploads/Guidelines/1724842648_ethical_guidelines_application_artificial_intelligence_biomed_rsrch_2023.pdf

cualquier aproximación meramente doméstica o regional. Se requiere una arquitectura normativa que combine principios éticos universales, mecanismos efectivos de rendición de cuentas y cooperación internacional.

El principio de precaución debe ocupar un lugar central. Ante la incertidumbre sobre los efectos de la IA en poblaciones vulnerables, la ausencia de certeza científica no puede invocarse para justificar la inacción⁹⁶. Ello implica suspender o limitar el uso de algoritmos opacos en ensayos clínicos hasta que existan evaluaciones independientes sobre su seguridad, equidad y validez externa. Aplicado estrictamente, impediría el despliegue de sistemas no auditados en estudios multicéntricos, garantizando que el imperativo hipocrático de “no hacer daño” prevalezca sobre la presión por acelerar la innovación⁹⁷.

Complementariamente, el principio de proporcionalidad exige calibrar los controles en función del riesgo y de la naturaleza del uso algorítmico. Un sistema de IA que incida en decisiones clínicas directas, como la asignación de dosis o la detección de eventos adversos graves, debe ser objeto de escrutinio intensivo. El reto consiste en imponer requisitos exigentes — transparencia, auditorías, validación independiente — sin sofocar la investigación responsable ni desincentivar desarrollos legítimamente beneficiosos⁹⁸.

La soberanía sanitaria y de datos constituye otro eje normativo. Los países del Sur global deben disponer de marcos jurídicos que impidan la extracción indiscriminada de datos clínicos y que otorguen a sus agencias reguladoras capacidad para evaluar tecnologías digitales de manera autónoma⁹⁹. Ello requiere leyes que prohíban la transferencia transfronteriza de datos sensibles sin autorización expresa, mecanismos de registro local de algoritmos y cooperación técnica internacional que refuerce las capacidades institucionales. La reciente posición de India al restringir colaboraciones que no garanticen una protección adecuada de los datos ilustra la viabilidad de este enfoque.

⁹⁶ COMMISSION OF THE EUROPEAN COMMUNITIES. *Communication on the precautionary principle*. Bruselas: Comisión Europea, 2000. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52000DC0001>

⁹⁷ WORLD MEDICAL ASSOCIATION. *Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects*. Fortaleza: WMA, 2013. Disponible en: <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>

⁹⁸ EUROPEAN PARLIAMENT AND COUNCIL. *Regulation (EU) 2024/1689 laying down harmonised rules on Artificial Intelligence (AI Act)*. Official Journal of the European Union, L, 12 julio 2024. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>

⁹⁹ INDIAN COUNCIL OF MEDICAL RESEARCH (ICMR). *Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare*. Nueva Delhi: ICMR, 2023. Disponible en: https://main.icmr.nic.in/sites/default/files/whats_new/ICMR_AI_Healthcare_Guidelines_2023.pdf

A su vez, la justicia epistémica debe incorporarse explícitamente en el diseño de la gobernanza global¹⁰⁰. Ello exige integrar a investigadores, comités de ética y comunidades locales en las decisiones sobre ensayos internacionales, reconociendo su papel no solo como ejecutores periféricos, sino como coproductores de conocimiento. La inclusión de representantes locales en órganos de aprobación, la coautoría en publicaciones y la cogestión de los datos y de los resultados constituyen medidas concretas para superar la tradicional asimetría Norte-Sur.

Un componente adicional es la transparencia algorítmica transnacional. Al igual que los registros públicos de ensayos clínicos, debería establecerse un registro internacional bajo los auspicios de la OMS u otra entidad multilateral donde se consignen las características esenciales de los sistemas de IA empleados en investigación biomédica: finalidad, datos de entrenamiento, validaciones realizadas y sesgos conocidos¹⁰¹. Este mecanismo permitiría un escrutinio global y daría a los pacientes información relevante para el consentimiento informado¹⁰².

La rendición de cuentas y los mecanismos de reparación son también imprescindibles. Las víctimas de daños ocasionados por algoritmos en ensayos internacionales suelen enfrentar barreras insalvables para acceder a la justicia. Resulta necesario crear esquemas globales que garanticen compensación, como fondos internacionales obligatorios financiados por los patrocinadores o mecanismos de arbitraje vinculantes administrados por organismos multilaterales. En paralelo, debe establecerse con claridad la responsabilidad jurídica: si un error algorítmico causa perjuicios, corresponde determinar si responde el desarrollador, el patrocinador del ensayo o ambos¹⁰³.

Finalmente, la colaboración multilateral debe traducirse en estándares compartidos y en la creación de un Código Internacional de Buenas Prácticas en Ensayos con IA, que complemente las normas de Buenas Prácticas Clínicas (ICH-GCP). Este código debería establecer criterios comunes sobre validación algorítmica, supervisión humana y protección de derechos fundamentales, con incentivos reputacionales y jurídicos para los actores que lo adopten.

¹⁰⁰ FRICKER, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press, 2007. ISBN 978-0-19-823790-7.

¹⁰¹ WORLD HEALTH ORGANIZATION. *Ethics and governance of artificial intelligence for health: WHO guidance*. Ginebra: World Health Organization, 2021. ISBN 978-92-4-002920-0. Disponible en: <https://www.who.int/publications/i/item/9789240029200>

¹⁰² UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. París: UNESCO, 2021. Disponible en: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>

¹⁰³ EUROPEAN COMMISSION. *Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*. Bruselas: Comisión Europea, 2022. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496>

En suma, avanzar hacia una gobernanza algorítmica transnacional exige articular precaución, proporcionalidad, soberanía, justicia epistémica, transparencia y rendición de cuentas. Solo mediante un marco coordinado de carácter universal será posible impedir que la inteligencia artificial profundice los desequilibrios históricos de la investigación global y, en su lugar, se convierta en una herramienta legítima de promoción de la salud y de los derechos humanos.

Conclusiones

La irrupción de la inteligencia artificial en la investigación farmacéutica global no puede entenderse como un mero avance tecnológico, sino como un fenómeno inserto en una economía política de la salud marcada por desigualdades históricas. El presente trabajo introduce y desarrolla la noción de neocolonialismo algorítmico aplicada a los ensayos clínicos internacionales, como categoría jurídica y crítica que permite describir con precisión cómo datos, cuerpos y territorios del Sur global pueden ser instrumentalizados bajo la opacidad de algoritmos diseñados y controlados desde el Norte. Esta aportación conceptual busca enriquecer el debate académico y ofrecer una herramienta analítica para identificar, visibilizar y enfrentar nuevas formas de dominación digital en el ámbito biomédico.

El examen de los marcos regulatorios vigentes muestra que el Derecho internacional aún carece de un instrumento vinculante capaz de afrontar estos desafíos. La fragmentación actual —reglamentos europeos con aplicación estricta dentro de su territorio, directrices éticas globales sin fuerza obligatoria, y legislaciones nacionales desiguales— revela no solo un vacío normativo, sino una crisis del Derecho internacional biomédico, que en su estado actual resulta incapaz de garantizar de manera universal la dignidad y la integridad de los sujetos de investigación. Esta carencia erosiona la legitimidad de un orden jurídico internacional que, para seguir siendo relevante, debe adaptarse a los retos de la era digital.

Ante este panorama, proponemos avanzar hacia una gobernanza algorítmica transnacional fundada en seis ejes: (i) el principio de precaución frente a tecnologías no auditadas; (ii) la proporcionalidad de las exigencias regulatorias según el nivel de riesgo; (iii) la soberanía sanitaria y de datos como expresión de igualdad entre Estados; (iv) la justicia epistémica, que reconozca a comunidades e investigadores locales como coproductores de conocimiento; (v) la transparencia algorítmica, mediante registros internacionales y auditorías independientes; y (vi) mecanismos efectivos de rendición de cuentas y reparación para las víctimas de ensayos. Estos pilares deben cristalizar en un instrumento jurídico internacional específico, ya sea un convenio vinculante o, al menos, una declaración de alto nivel que establezca estándares universales y evite la reproducción de dobles varas normativas. La constatación de estos vacíos impone la necesidad de un marco de gobernanza supranacional que supere el actual mosaico de normas y garantice coherencia entre innovación y derechos humanos.

La inteligencia artificial es una herramienta ambivalente: puede convertirse en un vector de dominación que profundice las brechas Norte-Sur, o en un instrumento emancipador capaz de detectar sesgos, reducir inequidades y ampliar el acceso a la salud. La dirección dependerá de la arquitectura jurídica y política que seamos capaces de construir.

La cuestión no es si habrá que regular la inteligencia artificial en la investigación farmacéutica, sino si lo haremos a tiempo para impedir que se convierta en el rostro digital de un viejo colonialismo. Permitirlo sería una renuncia del Derecho a su función más elemental: proteger la dignidad humana frente al poder. Regularla con rigor, en cambio, supondría afirmar la capacidad del Derecho internacional para responder a los desafíos del siglo XXI y garantizar que la revolución algorítmica se inscriba en la senda de la justicia sanitaria internacional. En última instancia, la regulación de la inteligencia artificial en la biomedicina no es un mero asunto técnico: es una prueba de la capacidad del Derecho internacional para reinventarse frente a los nuevos poderes. Si fracasa en ello, el futuro digital de la salud podría quedar regido por algoritmos sin justicia; si tiene éxito, convertirá la justicia sanitaria internacional en la brújula ética del siglo XXI.

Bibliografía

ABDULLAHI v. Pfizer, Inc. 562 F.3d 163 (2d Cir. 2009). *United States Court of Appeals for the Second Circuit*, 2009. Disponible en: <https://caselaw.findlaw.com/court/us-2nd-circuit/1310817.html>

ASOCIACIÓN MÉDICA MUNDIAL. *Declaración de Helsinki: Principios éticos para las investigaciones médicas en seres humanos*. Fortaleza: 64.ª Asamblea General de la AMM, 2013. Disponible en: <https://www.wma.net/es/polices-post/declaracion-de-helsinki-de-la-amm-principios-eticos-para-las-investigaciones-medicas-en-seres-humanos/>

BENJAMIN, Ruha; MULDOON, Julianna; et al. «Artificial Intelligence in the Colonial Matrix of Power». *Philosophy & Technology*, 2023, vol. 36, n.º 87.

COMISIÓN EUROPEA. *Comunicación sobre el principio de precaución*. Bruselas: Comisión Europea, 2000. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52000DC0001> [bmj.com](https://www.bmj.com)

COMISIÓN EUROPEA. *Propuesta de Directiva relativa a la responsabilidad por inteligencia artificial (AI Liability Directive)*. COM(2022) 496 final. Bruselas: 28 septiembre 2022. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496>

CORREA, Carlos María. *Intellectual Property and Public Health: The TRIPS Agreement and Patents on Medicines*. Ginebra: ICTSD, 2000. Disponible en: https://unctad.org/system/files/official-document/ictsd2000d2_en.pdf

COULDRY, Nick; MEJIAS, Ulises A. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford: Stanford University Press, 2019. ISBN 9781503609754.

COUNCIL FOR INTERNATIONAL ORGANIZATIONS OF MEDICAL SCIENCES (CIOMS). *International Ethical Guidelines for Health-Related Research Involving Humans*. Ginebra: CIOMS, 2016. ISBN 9789290360889. Disponible en: <https://cioms.ch/publications/product/international-ethical-guidelines-for-health-related-research-involving-humans/>

EUROPEAN MEDICINES AGENCY. *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle*. Amsterdam: EMA, 9 septiembre 2024. Documento EMA/CHMP/CVMP/83833/2023. Disponible en: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf

ESTADOS UNIDOS. *Alien Tort Statute*, 28 U.S.C. § 1350. Washington, D.C.: Office of the Law Revision Counsel, [ed. vigente].

FARMER, Paul. *Pathologies of Power: Health, Human Rights, and the New War on the Poor*. Berkeley: University of California Press, 2005. ISBN 9780520243262.

FRICKER, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press, 2007. ISBN 9780198237907.

GOLDRACRE, Ben. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. London: Fourth Estate, 2012. ISBN 9780007350742.

INDIAN COUNCIL OF MEDICAL RESEARCH (ICMR). *Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare*. Nueva Delhi: ICMR, 2023. Disponible en: https://www.icmr.gov.in/icmrobject/uploads/Guidelines/1724842648_ethical_guidelines_application_artificial_intelligence_biomed_rsrch_2023.pdf

INTERNATIONAL COUNCIL FOR HARMONISATION (ICH). *Guideline for Good Clinical Practice E6(R2)*. Ginebra: ICH, 2016. Disponible en: <https://ichgcp.net/>

KAMINSKI, Margot E.; MALGIERI, Gianclaudio. «Algorithmic impact assessments under the GDPR: producing multi-layered explanations». *International Data Privacy Law*, 2021, vol. 11, n.º 2, pp. 125-144. DOI: 10.1093/idpl/ipaa020. [OUP Academic](#)

NACIONES UNIDAS. *Pacto Internacional de Derechos Económicos, Sociales y Culturales*. Resolución 2200 A (XXI), 16 dic 1966; en vigor 3 ene 1976. *United Nations Treaty Series*, vol. 993, p. 3.

NACIONES UNIDAS. *Observación General n.º 14: El derecho al disfrute del más alto nivel posible de salud (art. 12 del PIDESC)*. Ginebra: Comité DESC, 2000.

ORGANIZACIÓN MUNDIAL DE LA SALUD (OMS). *Ética y gobernanza de la inteligencia artificial para la salud*. Ginebra: OMS, 2021. ISBN 978-92-4-002920-0. Disponible en: <https://www.who.int/publications/i/item/9789240029200> [Nigerians to sue US drug company over meningitis treatment - PMC https://pmc.ncbi.nlm.nih.gov/articles/PMC1172920/](https://pmc.ncbi.nlm.nih.gov/articles/PMC1172920/)

PETRYNA, Adriana. *When Experiments Travel: Clinical Trials and the Global Search for Human Subjects*. Princeton: Princeton University Press, 2009, vol. 5, n.º 2. Disponible en: <https://scispace.com/pdf/when-experiments-travel-clinical-trials-and-the-global-1pkd95gsut.pdf>

RAJAN, Kaushik Sunder. *Biocapital: The Constitution of Postgenomic Life*. Durham: Duke University Press, 2006. ISBN 9780822337349.

SOUTH CENTRE (de Menezes, H.Z.). *Negotiating Health and Autonomy: Data Exclusivity, and the Limits to TRIPS Harmonization*. Research Paper 204. Ginebra: South Centre, 2024. Disponible en: https://www.southcentre.int/wp-content/uploads/2024/07/RP204_Negotiating-Health-and-Autonomy_EN.pdf

SUPREME COURT OF INDIA. *Swasthya Adhikar Manch v. Union of India & Ors.*, Writ Petition (Civil) No. 33 of 2012, sentencia de 3 ene 2013. Disponible en: <https://indiankanoon.org/doc/11325454/>

UNESCO. *Recomendación sobre la Ética de la Inteligencia Artificial*. París: UNESCO, 2021.

UNESCO. *Declaración Universal sobre Bioética y Derechos Humanos*. París: UNESCO, 2005.

UNIÓN EUROPEA. *Reglamento (UE) 2016/679 (RGPD)*. DOUE L 119, 4 may 2016. Disponible en: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

UNIÓN EUROPEA. *Reglamento (UE) 536/2014, sobre ensayos clínicos de medicamentos de uso humano*. DOUE L 158, 27 may 2014. Disponible en: <https://eur-lex.europa.eu/eli/reg/2014/536/oj>

UNIÓN EUROPEA (Parlamento y Consejo). *Reglamento (UE) 2024/1689 por el que se establecen normas armonizadas en materia de inteligencia artificial (AI Act)*. DOUE L 237, 12 jul 2024. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>

VEALE, Michael; ZUIDERVEEN BORGESIJUS, Frederik. *Demystifying the Draft EU Artificial Intelligence Act*. *Computer Law Review International*, 2021, vol. 22, n.º 4, pp. 97-112. DOI: 10.9785/crl-2021-220402. <https://discovery.ucl.ac.uk/id/eprint/10131593/>

WISE, J. «Pfizer accused of testing new drug without ethical approval». *BMJ*, 2001, vol. 322, n.º 7280, p. 194. Disponible en: <https://www.bmj.com/content/322/7280/194.1>

YAHYA, Maryam. «Polio vaccines—“no thank you!” barriers to polio eradication in northern Nigeria». *African Affairs*, 2007, vol. 106, n.º 423, pp. 185-204. DOI: 10.1093/afraf/adm016.

Pablo Ceballos. Regulación IA y riesgos democráticos: Un crítico análisis de los Derechos Humanos en El Salvador

Pablo Ceballos

Universitat de València

[ACCEDE A LA GRABACIÓN](#)

Resumen

La reciente aprobación en El Salvador de la Ley de Fomento de la Inteligencia Artificial refleja un esfuerzo por integrar la innovación tecnológica en la gobernanza y el desarrollo económico. La normativa contempla la creación de la Agencia Nacional de Inteligencia Artificial (ANIA) y principios éticos orientados a la equidad, transparencia y responsabilidad. Sin embargo, su carácter inicial deja vacíos significativos en la regulación de riesgos y en la protección de los derechos humanos. Este trabajo analiza los alcances y limitaciones de la ley, señalando tanto sus potenciales beneficios como los riesgos de instrumentalización política y retroceso democrático.

Palabras clave

Inteligencia Artificial, Gobernanza, El Salvador, Derechos Humanos, Regulación Tecnológica, Democracia, Innovación Legal.

Abstract

The recent approval of El Salvador's Artificial Intelligence Promotion Law reflects an effort to integrate technological innovation into governance and economic development. The regulation includes the creation of the National Artificial Intelligence Agency (ANIA) and ethical principles based on equity, transparency, and responsibility. However, as a first-generation law, it presents significant gaps in risk regulation and human rights protection. This paper examines the scope and limitations of the legislation, highlighting both its potential benefits for public services and the economy, as well as the risks of political instrumentalization and possible democratic setbacks.

Keywords

Artificial Intelligence, Governance, El Salvador, Human Rights, Technological Regulation, Democracy, Legal Innovation.

Introducción

Los avances tecnológicos y computacionales han permitido crear una herramienta como la Inteligencia Artificial, propia de distopías literarias pero que ahora protagoniza la realidad de todos en un período relativamente breve de adaptación. Esta tecnología ha permitido establecer la mayor revolución tecnológica que se ha vivido a lo largo de la historia de los seres humanos pudiendo incluso llegar, en numerosas ocasiones, a hacernos cuestionar qué es lo que realmente nos define como humanos y nos diferencia del resto de especies y elementos existentes de la Tierra. Esta transición tecnológica supone un salto cualitativo mayor al que se vio con los algoritmos, ya que no se limitan a hacer una única acción a partir de instrucciones precisas y secuenciales. La IA es capaz de reproducir el funcionamiento de nuestro cerebro sin requerir la existencia de una programación previa para el descubrimiento de patrones, funciones de aprendizaje, de pensamiento o de razonamiento. La aparición e integración de estos nuevos modelos tecnológicos dan lugar a la necesidad de regulación normativa y discusión legislativa que afecta tanto a nivel global como regional, pero que trataremos de centrarnos en la realidad que atiende a la República de El Salvador.

Al igual que ya se está realizando y se ha realizado en numerosos Estados, el Estado de El Salvador decidió aprobar el pasado 3 de marzo de 2025 en su Asamblea Legislativa una ley de fomento de la Inteligencia Artificial y de las Tecnologías. Esta decisión podría llegar a considerarse coherente y necesaria dada la rápida implementación de estas nuevas inteligencias en la vida común de las personas. No obstante, resulta igualmente necesaria la realización de un arduo análisis que refleje la aplicación de la actual innovación legal en un contexto en el que se producen constantes violaciones de derechos humanos en el territorio desde la llegada de Nayib Bukele al gobierno en 2019 y agravada desde la instauración del estado de excepción en el 2022. Esto se ha visto reflejado en la suspensión de los derechos humanos de la población con el objetivo de preservar un interés de seguridad nacional que había sido alterado por la situación prolongada de violencia provocada por las “maras” o pandillas.

En definitiva, este estudio pretende aportar luz sobre la situación de las graves vulneraciones de derechos humanos que afecta a El Salvador que, a pesar de su magnitud, sigue siendo invisibilizada en muchos espacios. El Salvador se encuentra en una situación de aparente impunidad, invisibilizada por el control mediático nacional e internacional del uso de las redes sociales y que, a través de una aplicación indebida de esta reciente legislación podría llegar a instrumentalizar estas continuadas violaciones sistemáticas y aumentar los riesgos de los neuroderechos, así como el derecho a la identidad, el libre albedrío, la privacidad, el acceso equitativo a las tecnologías de mejora y los sesgos de algoritmos respecto a las voces opositoras. Redirigiendo de manera encubierta los pensamientos de la población mediante la influencia sobre el discurso público y sobre los medios de comunicación.

1. Una nueva era sin suficiente regulación armonizada

No es una novedad que el derecho a menudo se sitúe por detrás de las innovaciones tecnológicas, pues acaba resultando enormemente complicado poder normativizar algo cuando ni siquiera se conoce el funcionamiento y el impacto sobre la vida de las personas. Aunque en el mundo cibernético no se puede restringir la acción de los Estados a una mera libertad de desarrollo ya que, si no se pueden restringir ciertas limitaciones al incontrolado desarrollo tecnológico, se podrían llegar a multiplicar los riesgos que le acompañan. Las posibilidades y dificultades que entrelazan la presente evolución digital, las cuales se ven exponencialmente multiplicadas con el crecimiento y difusión de la inteligencia artificial en el ámbito político, no solo suponen una extrema actualidad, sino que también requieren una especial atención política y jurídica. Pues, en caso de no existir un adecuado acompañamiento paralelo de la progresión tecnológica mediante un control legal, podría conllevar a un desbordamiento imposible de sostener al mismo tiempo que se produjese un retroceso político y social¹⁰⁴, como ya se puede observar en una pluralidad de Estados.

Los desafíos que presenta la inteligencia artificial (IA) exigen cerrar brechas significativas y avanzar hacia una regulación armonizada, construida a partir de acuerdos nacionales y regionales. Solo de esta forma será posible impulsar políticas integrales y coherentes que equilibren el fomento y la regulación de los marcos de desarrollo de la IA en América Latina. Este proceso requiere, además, del respaldo institucional de líderes y autoridades, así como de una adecuada asignación de recursos, que garanticen un desarrollo seguro, sostenible y con relevancia estratégica de estas tecnologías.

Entre este entramado jurídico existe el debate sobre la creación de un marco de normas de IA respecto a cómo regular una cuestión tan heterogénea que transide desde sistemas de reconocimiento facial a vehículos autónomos hasta sistemas predictivos y generativos, herramientas, funcionalidades, técnicas y diferentes estados de autonomía de la IA que con frecuencia plantean a los expertos si la respuesta sería establecer una regulación general e integral.

La ausencia de regulación armonizada no implica únicamente la creación de una normativización extensiva del funcionamiento y control de los riesgos y peligros de la IA, sino que más bien consista en una regulación eficaz y eficiente efectuada por medio del conocimiento necesario de la profundidad de esta tecnología por parte de los políticos y

¹⁰⁴ CASTELLANOS CLARAMUNT, J., “La Inteligencia Artificial en el contexto jurídico: El derecho como garante del desarrollo democrático”, *Diálogos Jurídicos*, 2024, pp. 91 – 92.

legisladores que participen y promuevan la elaboración de normas que controlan la IA. Por una parte, existen con frecuencia riesgos y mitos exagerados de las Inteligencias Artificiales, al igual que no se llegan a cubrir los peligros que realmente pueden preocupar. No consiste en una cuestión de regulación invidente sino en categorizar los sistemas de acuerdo con su riesgo y el establecimiento del cumplimiento de obligaciones de estos. Una intervención regulatoria temprana sobre la prevención de daños permite reducir los posibles impactos negativos paralelamente a la reducción proporcional de los costes de reparación. Esta tendencia ya fue expresada por los países miembros del G7 (Canadá, Francia, Alemania, Italia, Japón, Reino Unido, Estados Unidos y la UE) en cuanto a la intención de unificar la regulación de IA en base a políticas públicas en la materia del riesgo, alcanzando así un acuerdo sobre los Principios Rectores Internacionales sobre la IA y sobre un Código de conducta para desarrolladores.¹⁰⁵

A diferencia de instituciones supranacionales como la Unión Europea, existen grandes vacíos regulatorios a nivel global desde una perspectiva armonizada. Pues, países como Estados Unidos o China, establecen sus propias regulaciones que restringen o liberan según como consideren el beneficio tecnológico que pudiese aportar. En una cuestión de revolución tecnológica como se trata, es necesario un sentido y una voluntad común para que persista el avance tecnológico de manera conjunta, pero sin afectar a derechos y bienes jurídicos tan importantes y necesarios como los derechos humanos. En un escenario de innovación y guerra fría tecnológica, los derechos humanos se mantienen en un plano subalterno, alejado de los intereses que comporta el desarrollo de la IA. Aunque, es parte de la potestad y competencia regulatoria de los Estados de enfrentarse regulatoriamente a las empresas que tejen la red de dominio y control de la IA. Pues, ya no se trata de códigos de conducta y principios rectores de funcionamiento, sino del establecimiento de límites infranqueables que impidan la vulneración de los derechos y que sean impermeables a las consecuencias de la IA debido a una respuesta tardía regulatoria.

En el caso europeo, cuando se presentó el plan estratégico de la Comisión Europea en 2019, se planteó la regulación digital como un elemento clave de la legislación de la UE para el período 2019-2023. Hasta el momento, solo existían escasas normativas sobre cuestiones digitales, como la Directiva 95/46/CE sobre datos personales, la directiva que derivó de esta sobre privacidad electrónica, la Directiva 2000/31/CE sobre comercio electrónico y la Directiva 2013/37/UE sobre información del sector público. Frente a esta propuesta normativa, el objeto del legislador europeo, no se enfocaba primordialmente en la protección de los derechos fundamentales, sino más bien en fomentar el desarrollo de la IA en una situación de debilidad

¹⁰⁵ CORVARÁN, J. G. & CARRO, M.V., *Op. Cit.*, pp. 92-95

industrial del sector en Europa. Para el diseño de estas leyes de innovación legal se ha pretendido utilizar el recurso tradicional a la responsabilidad extracontractual que se ha adoptado negativamente a la situación de complejidad tecnológica, los operadores globales con amplias capacidades financieras, la reducción de los daños y, la confianza de la estimulación de las nuevas tecnologías, como la denominada IA fiable, así como la garantía de establecer entornos seguros tecnológicos en lugar de remedios de compensación de consecuencias desastrosas.¹⁰⁶

En la región latinoamericana y especialmente, la centroamericana, existe un crecimiento paulatino regulatorio sobre el interés de la IA. Aunque el crecimiento que mantienen estos países acaba situándose en una posición inferior con respecto al resto de países latinoamericanos y, más acentuada a nivel global. Esta cuestión no acaba siendo limitante siempre y cuando se ejecuten los debidos procesos normativos que protejan las carencias que otros países no contemplaron en los enfoques basados en el riesgo de una legislación de primera generación. Aunque, la realidad no está siendo como debería efectuarse.

En base al informe que presentó el Índice Latinoamericano de IA revela un panorama desigual entre los 19 países evaluados, pues únicamente Chile, Uruguay y Brasil destacan por su sólido y equilibrado desempeño en investigación, desarrollo y adopción de la IA. Les siguen México, Argentina y Colombia, que pese a enfrentar limitaciones en infraestructuras y en políticas de apoyo -los que el índice denomina Factores Habilitantes-, mantienen un impulso significativo en la generación de conocimiento en torno a esta tecnología. En contraste, la mayoría de los países restantes ocupan posiciones más rezagadas, en particular aquellos de Centroamérica y El Caribe, incluyendo a la República Dominicana, Cuba, Ecuador, Perú, Costa Rica, Panamá, Venezuela, Bolivia, El Salvador, Honduras y Guatemala.¹⁰⁷

Pese al desarrollo en IA durante los últimos ocho años en los países de América Latina y el Caribe, no se ha alcanzado en ningún caso los niveles evidenciados en el Norte Global. Esto muestra una brecha que se mantiene en la región y que refiere a una escasez de talento e investigación en esta materia. En el territorio latinoamericano se observa un promedio de alfabetización en IA relativamente menor al de los países del Norte Global por lo que las políticas de alfabetización y el fomento en la implementación de Inteligencia Artificial podrían suponer una gran oportunidad para garantizar nuevas oportunidades laborales en la región. La encrucijada persiste en la retención de talento y en la formación de este. Puesto que, desde el

¹⁰⁶ MANTELERO, A., “El Reglamento de inteligencia artificial: la respuesta del legislador europeo a los retos de la inteligencia artificial”, en *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Editorial Aranzadi, 2024, pp. 54 – 58.

¹⁰⁷ CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL (CENIA), *Índice Latinoamericano de Inteligencia Artificial*, 2024, pp. 8 – 31

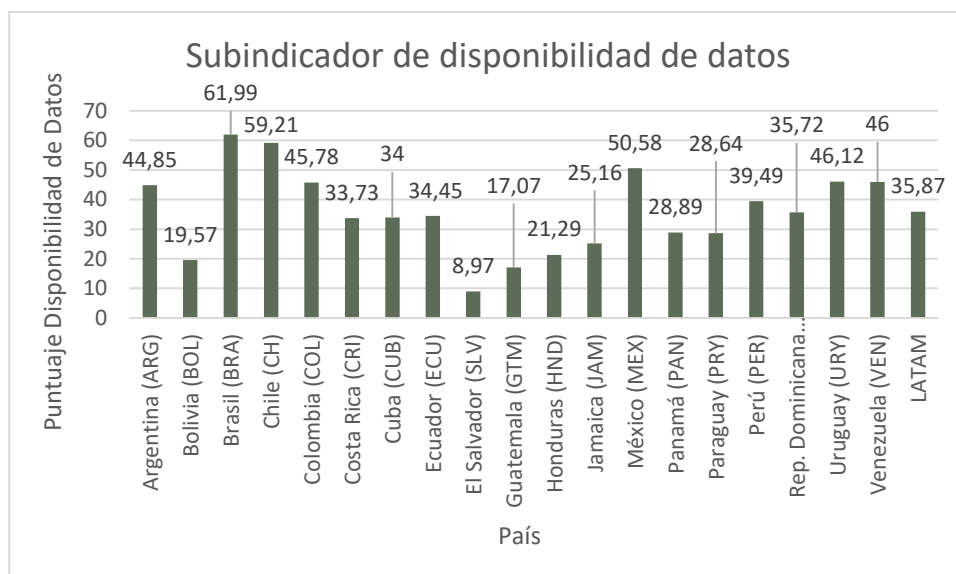
2019 se ha detectado una tendencia permanente en todos los países -excepto Costa Rica y Uruguay- que han perdido más talento del que han atraído, dado que en términos migratorios los países no se perfilan como atractores de talento IA. También, se observa unos patrones persistentes en la brecha de género que se evidencian en ingeniería de IA y se replican en el desarrollo de competencia de alfabetización en esta materia. La participación de mujeres en este sector es del 27%, mientras que en cuanto a la alfabetización se encuentra en el 22%. Este preocupante panorama refleja la brecha contraintuitiva que se establece en cuestiones de género. Además, resulta también un reto la formación de habilidades básicas, tales como el pensamiento crítico, pensamiento computacional y relacionados con los sistemas educativos respecto al acceso justo y equitativo a la tecnología.¹⁰⁸

El acceso y disponibilidad a datos abiertos y fiables, junto con la protección de los datos personales, constituye un pilar esencial para el desarrollo adecuado de la inteligencia artificial. Sin conjuntos de datos de calidad y libre disponibilidad, simplemente no existiría la materia prima necesaria para entrenar los algoritmos y modelos de aprendizaje automático. Para evaluar esta dimensión de los datos, resulta clave el Barómetro de Datos¹⁰⁹ que se estructura bajo los subindicadores que analizan la disponibilidad, la capacidad, la gobernanza y el uso e impacto para generar un ecosistema saludable de IA.

El subindicador de disponibilidad mide la claridad, accesibilidad y facilidad de procesamiento de los datos públicos para uso abierto. Más allá de un aspecto meramente técnico, este indicador expone la profunda desigualdad en el acceso a los datos, condición indispensable para un aprovechamiento equitativo de la inteligencia artificial. La disparidad evidenciada muestra cómo el acceso restringido a la información limita las oportunidades de crecimiento tecnológico y amplía la brecha entre países con mayores recursos y aquellos rezagados. A nivel regional, el promedio se sitúa en 35,87 puntos, con Brasil (61,99), Chile (59,21) y México (50,58) como líderes relativos. Mientras que, El Salvador (8,97) ocupa la última posición, reflejando no solo una carencia estructural de datos accesibles, sino también una fragilidad institucional que dificulta sentar las bases de un ecosistema de IA inclusivo y democrático.

¹⁰⁸ CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL (CENIA), *Op. Cit.*, pp. 37-39

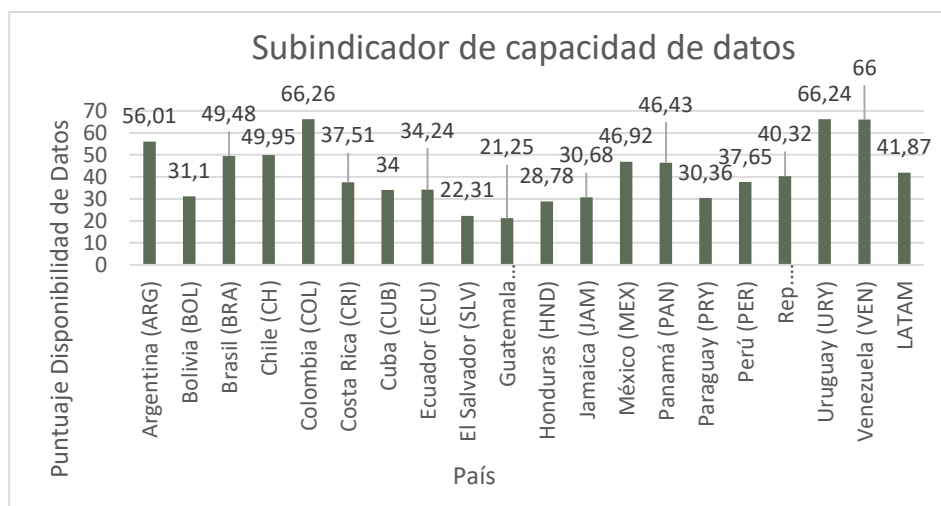
¹⁰⁹ Este barómetro es el único indicador sistemático elaborado por el Barómetro Global de Datos. Este instrumento recopila y analiza la información abierta disponible en cada país con el propósito de medir la provisión de datos esenciales para la toma de decisiones gubernamentales en múltiples áreas estratégicas, bajo el parámetro de los datos esenciales para la toma de decisiones gubernamentales en diferentes áreas de acción.



Fuente: Gráfico propio - Datos: ILIA 2024/ Global Barometer's

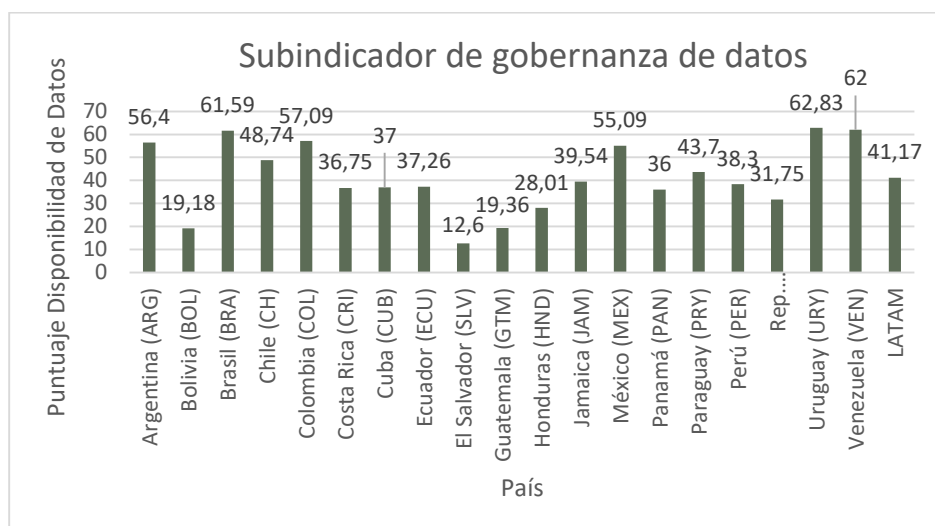
El subindicador de capacidades evalúa la eficacia con que los países recopilan, procesan, utilizan y comparten datos, considerando factores como conectividad, formación profesional e instituciones que los respalden. Con un promedio regional de 41,87 puntos, los resultados revelan un escenario desigual, aunque países como Colombia (66,26), Uruguay (66,24) y Venezuela (66,00) demuestran un nivel avanzado de preparación, alrededor del 50% de los países de la región se mantiene por debajo de este umbral, lo que pone en evidencia una brecha estructural en la capacidad de aprovechar el potencial de los datos. En este contexto, El Salvador (22,31) se sitúa apenas por encima de Guatemala (21,26), lo que subraya la inadaptación tecnológica del país hasta 2024 y la falta de condiciones institucionales y formativas para integrarse de manera efectiva en la transformación digital.¹¹⁰

¹¹⁰ CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL (CENIA), *Op. Cit.*, pp. 77-78



Fuente: Gráfico propio - Datos: ILIA 2024/ Global Barometer's

El subindicador de gobernanza refleja la capacidad de los países para establecer reglas, procesos y estructuras que garanticen derechos fundamentales vinculados a los datos, así como el acceso a información veraz y completa, transparencia en su gestión y protección de la privacidad personal. Sin embargo, los resultados muestran un panorama fragmentado. El promedio regional se ubica en 41,17 puntos, pero con contrastes significativos. Países como Uruguay, Venezuela y Brasil superan los 60 puntos, mientras que un total de 11 países permanecen por debajo de ese umbral. Este constante desequilibrio evidencia la desigualdad estructural en la gobernanza de datos en América Latina, donde coexisten marcos regulatorios relativamente sólidos con contextos de debilidad institucional y normas incipientes.

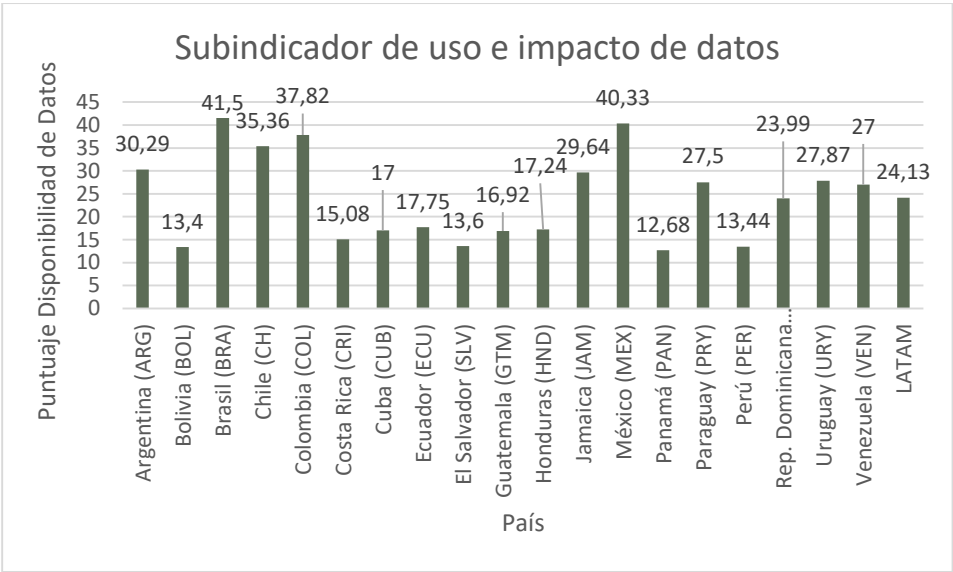


Fuente: Gráfico propio - Datos: ILIA 2024/ Global Barometer's

El caso de El Salvador (12,60), que ocupa el último lugar, resulta especialmente preocupante. Le siguen de cerca Bolivia (19,18) y Guatemala (19,36), evidenciando no solo la ausencia de políticas específicas en la protección y gestión de datos, sino también una vulnerabilidad persistente frente a riesgos como la manipulación informativa, la falta de transparencia

gubernamental y la precariedad en la garantía de derechos digitales. Así, este subindicador nos muestra que, en gran parte de la región, los marcos normativos avanzan de forma lenta, desigual y reactiva, comprometiendo la posibilidad de construir un ecosistema de datos que sirva de base confiable y democrática para el desarrollo de la inteligencia artificial.

Por último, el subindicador de uso e impacto examina cómo se aplican los datos en la práctica, identificando tanto los casos de uso más representativos como los sectores sociales o productivos que muestran mayor interés en aprovecharlos. Los resultados evidencian nuevamente una marcada heterogeneidad regional, situándose el promedio regional en 24,13 puntos, lo que revela un nivel limitado de madurez en la incorporación de los datos como motor de innovación y desarrollo.



Fuente: Gráfico propio - Datos: ILIA 2024/ Global Barometer’s

La situación es especialmente crítica en la cuenca del Caribe, donde se concentran las puntuaciones más bajas. El Salvador (13,60) se ubica ligeramente por encima de Perú (13,44), Bolivia (13,40) y Panamá (12,68), lo que confirma la fragilidad de estos países en la capacidad de transformar datos en soluciones de impacto real. En este escenario, no sólo se muestra el bajo aprovechamiento de los datos, sino también la ausencia de políticas e infraestructuras que fomenten su uso estratégico en beneficio de la ciudadanía y de la competitividad económica.¹¹¹

Por esta razón, la débil gestión de los parámetros analizados anteriormente evidencia la necesidad urgente de diseñar una normativa adaptada que facilite la incorporación gradual de

¹¹¹ CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL (CENIA), *Op. Cit.*, pp. 79-80

los Estados latinoamericanos y caribeños al ecosistema de la inteligencia artificial. En la actualidad, la mayoría de los países apenas cuentan con planes estratégicos preliminares y con un conjunto de recomendaciones éticas que, aunque valiosas, resultan insuficientes para responder a los desafíos prácticos de la IA. Si bien casi todos disponen de leyes de datos relativamente recientes, el panorama normativo aún está fragmentado. Sin embargo, proyectos de ley más avanzados en países como México y Brasil, denotan una mayor profundización regulatoria y llegar a considerarse precedentes relevantes para la región.

Asimismo, para llevar a cabo el correcto desarrollo de IA en los diferentes ámbitos (empresarial, gubernamental y familiar) se precisa de la unificación de esfuerzos y recursos mediante un enfoque colaborativo, de manera que dicha tecnología emergente cuente con los elementos requeridos para una correcta adaptación al contexto global. Es preciso mencionar que, el 16% de la región en el 2022 ya habían adoptado el uso de la IA, pero en contraste con Centroamérica (Guatemala, El Salvador, Costa Rica y Panamá) este promedio se redujo al 7.25%, relegando la posición de las pequeñas y medianas empresas salvadoreñas a una estimación aproximada al 8% en el desarrollo de la IA. Observándose una destinación de recursos para IA a una posición rezagada.¹¹²

Lo cierto es que, para la implementación de la IA en la vida pública y privada, así como la muestra de un interés sobre la regulación, existe una interpretación bicéfala entre los Estados latinoamericanos. Por una parte, no es casualidad que México, Chile, Brasil o Argentina, se hayan encontrado unas tendencias de integración superiores al resto de países latinoamericanos. Pues, a pesar de que existan deficiencias en las dinámicas democráticas, suelen contar con instituciones mucho más sólidas que las que se han visto en otros países.

Como norma general, la presencia de instituciones democratizadas favorece la implementación y regulación de nuevas tecnologías. Aunque, así como también hemos visto, el carácter regulatorio de países autoritarios como China, establece una perspectiva diferenciada sobre la capacidad regulatoria y los intereses que se muestran sobre qué regular. El problema reside en países como El Salvador, que sin tener unas instituciones sólidas y progresivamente, un régimen autoritario más consolidado, tenga la intención de regular sobre Inteligencia Artificial, sin determinar cuál será la capacidad de impacto y de acción en esta materia.

No obstante, aunque los datos en 2024 apuntaban a la continuada situación rezagada de El Salvador en cuestiones normativas de inteligencia artificial, la creación de la presente norma

¹¹² GARCÍA CABRERA, B.A., “Inteligencia Artificial y economía”, *Revista Ciencias Económicas*, 1 (1), San Salvador, El Salvador, 2023, pp. 31-36.

establece un gran salto en el proceso de iniciativa regulatoria. En este instante, únicamente Perú y El Salvador cuentan con una legislación sobre IA aprobada, planteando la situación en la que se debe continuar exigiendo una prolongada discusión de esta materia -dada la reducida madurez legal-, debiéndose evaluar en el tiempo los posibles riesgos y peligros de su desarrollo e impacto legal.

2. Ley de Inteligencia Artificial de El Salvador

Debido a la necesidad incipiente del país centroamericano, la Asamblea Legislativa aprobó el 25 de febrero de 2025, la Ley de Fomento de la Inteligencia Artificial y Tecnologías, aprobada con 57 votos a favor de los 60 posibles.¹¹³ A simple vista ya muestra una limitada negociación legislativa, representada por su amplia mayoría parlamentaria que no debe haber sufrido graves obstáculos para la aprobación de este decreto. La diputada opositora Claudia Ortiz calificó el proyecto de “deficiente” y reclamó un debate más profundo sobre las implicaciones éticas, económicas y laborales. Puesto que, la aprobación de esta ley limita la participación de actores clave como académicos, ciudadanos o sindicatos en la definición de la norma.¹¹⁴

El hecho de portar los intereses legislativos en materia de Inteligencia Artificial a la Asamblea Legislativa de El Salvador es sumamente positivo para la creación normativa de estos parámetros, pero como bien apuntábamos antes, la regulación de legislación de primera generación también debe exigir una amplitud de voces que permitan una correcta deliberación democrática.

Por su parte, el objetivo de esta ley centra en impulsar el desarrollo, la investigación y la aplicación de la inteligencia artificial en El Salvador (art. 1 LFIAT), fomentando así la innovación y asegurando su uso responsable para la contribución en el avance tecnológico y el crecimiento económico del país. Esta nueva ley aplicaría a todas las personas naturales o jurídicas involucradas en la investigación, desarrollo y la aplicación de la inteligencia artificial, así como también a quienes almacenen, recopilen y procesen datos para estas actividades (art. 3 LFIAT).

En esta norma, se establecen una serie de principios éticos reflejados en la ejecución de las actividades que se contemplen en la ley mediante el respeto de los principios de equidad,

¹¹³ Asamblea Legislativa de la República de El Salvador, “Ley de Inteligencia Artificial ayudará a mejorar los servicios públicos e impulsar la economía”, 2025.

Disponible en: <https://www.asamblea.gob.sv/node/13492#:~:text=La> (Consultado el 22 de septiembre de 2025)

¹¹⁴ LÓPEZ, G., “Diputada Ortiz pide un debate amplio sobre la propuesta de ley de fomento de la inteligencia artificial “en términos éticos y económicos”, LaNoticiaSV, 2025.

Disponible en: <https://lanoticiasv.com/diputada-ortiz-pide-un-debate-amplio-sobre-la-propuesta-de-ley-de-fomento-de-la-inteligencia-artificial-en-terminos-eticos-y-economicos/> (Consultado el 24 de septiembre de 2025)

transparencia, responsabilidad, consentimiento informado, minimización de datos, inclusión y no discriminación (art. 5 LFIAT). Con el objeto de garantizar una correcta implementación de la ley, se articula mediante la misma la Agencia Nacional de Inteligencia Artificial (ANIA) que mantendrá una autonomía funcional y técnica, actuando como una institución desconcentrada y adscrita a la Presidencia de la República. A esta se le atribuyen principales funciones como la supervisión del cumplimiento de la ley y la gestión del Marco Integral de Evaluación de Riesgos para la IA y tecnologías similares para equilibrar la innovación tecnológica con la seguridad pública y el bienestar social (art. 17 LFIAT). Aunque las principales atribuciones se centrarán en la coordinación y supervisión del cumplimiento de las obligaciones, establecer programas de asistencia técnica y capacitaciones a la población, coordinar las actividades estatales sobre el fomento de la IA, conformar y administrar el Registro Nacional de Desarrollo, Innovación y Aplicación de IA, colaborar con el Ministerio de Educación, Ciencia y Tecnología en la incorporación de educación en inteligencia artificial en el sistema educativo, brindar asistencia técnica a las instituciones públicas, realizar una rendición de cuentas anual y, fomentar las asociaciones público-privadas que impartan formación práctica con herramientas y tecnologías estándar de la industria (art. 8 LFIAT).

En cuanto a la protección de datos personales y seguridad, esta ley tendrá un carácter de aplicación especial respecto otras leyes, sin perjuicio de lo que establece la Ley de Ciberseguridad y Seguridad de la Información y la Ley para la Protección de Datos Personales (art. 28 LFIAT), así como las normativas de seguridad emitidas y bajo la supervisión de la Agencia Nacional de Inteligencia Artificial (ANIA) y la Agencia de Ciberseguridad del Estado (ACE) (art. 22 LFIAT).

La protección de la propiedad intelectual internacional aplicables a la IA será facilitada por el Estado salvadoreño mediante el registro de las innovaciones de los desarrolladores en tratados globales como el Tratado de Cooperación en Materia de Patentes (PCT) (art. 24 LFIAT). Reconociendo la importancia de la existencia de una inteligencia artificial de código abierto y de libre flujo de conocimientos para el avance tecnológico, un mayor crecimiento económico y un desarrollo nacional. Se prohibirá así la imposición restrictiva al uso de manera arbitraria, modificación, desarrollo o integración de modelos, conjuntos de datos, ponderaciones, algoritmos o software de IA, siempre que se cumpla la legislación salvadoreña y los acuerdos

internacionales, garantizando la colaboración y acceso a diferentes innovaciones tecnológicas. (art. 26 LFIAT).¹¹⁵

A grandes rasgos, el texto normativo parece cumplir con las exigencias mínimas de regulación de esta materia, pero aun así existen deficiencias sustanciales que afectan tanto en su contenido formal como material. Si bien no existen muchos países que cuenten con una legislación sólida en esta materia, produce una desventaja comparativa en su elaboración normativa. Aunque, lo cierto es que ya existen precedentes en los últimos años que demuestran los esfuerzos de regulación de la IA, concentrándose en recomendaciones éticas, planes estratégicos y actualizaciones de los ecosistemas de protección de datos personales. Véase el caso de México, Chile, Uruguay, Brasil, Colombia o Argentina que mostraron un avance progresivo en las estrategias para establecer marcos regulatorios para la IA y en los que coexistiese guías éticas que orientasen su desarrollo responsable.¹¹⁶

En cuestiones de forma, la ley no cuenta con una amplia extensión que defina correctamente las obligaciones del Estado y los peligros que pudiese conllevar el desarrollo y uso de la inteligencia artificial, al igual que tampoco cómo el Estado podría afrontarlos. A nivel material, muestra una serie de definiciones sobre la conceptualización de los elementos referidos a la IA -con un diseño similar a la legislación peruana nº 31814¹¹⁷- y que más allá de establecer su significado no muestra la relevancia en el contenido de la norma, sin establecer efectos tangibles en sus limitaciones.

Por su parte, el 18 de julio de 2025, decretaron una reforma a diferentes artículos de la ley publicada en marzo del mismo año con la intención de fortalecer la estructura administrativa de la Agencia Nacional de Inteligencia Artificial y dotarla de un mayor grado de competencias para realizar sus funciones. Ampliando su autonomía más allá de considerarla únicamente funcional y técnica, dotándola de carácter administrativo y presupuestario (art. 7 LFIAT). Se extendieron sus atribuciones y actividades, especificando el origen patrimonial y presupuestario de la institución (art. 8-A. LFIAT). Se profundizó en las funciones del director ejecutivo de la ANIA y se desarrollaron las competencias de la estructura organizativa en su reglamento interno (art.9 LFIAT). Al igual que la creación de un reglamento interno de gestión

¹¹⁵ Asamblea Legislativa de la República de El Salvador, *Ley de fomento a inteligencia artificial y tecnologías*, Decreto nº234, 3 de marzo de 2025.

¹¹⁶ CORVARÁN, J. G. & CARRO, M.V., *Op. Cit.*, pp. 100-109.

¹¹⁷ Congreso de la República de Perú, *Ley que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país*, Ley nº31814, 2023.

Disponible en: <https://cdn.www.gob.pe/uploads/document/file/5038703/ley-que-promueve-el-uso-de-la-inteligencia-artificial-en-fav-ley-n-31814.pdf?v=1692895308>

de la ANIA que ampliase los elementos que lo relacionan con el directivo ejecutivo (art. 11 LFIAT).¹¹⁸

La figura de la ANIA es de suma importancia, dado que este organismo va a poder crear nuevas normativas técnicas. Tales como, la normativa de Registro del Nacional de Desarrollo, Innovación y Aplicación de IA, que entraría en vigor en un plazo máximo de noventa días a partir de la vigencia de la Ley y que abrió el procedimiento y los requisitos sobre el registro obligatorio para la inscripción de los modelos y sistemas que trabajan de manera privada (art.27 LFIAT). La ANIA estableció al mismo tiempo, un Marco Integral de Evaluación de Riesgos para la IA para aquellas personas físicas o jurídicas que trabajasen con datos confidenciales, reservados o personales (art. 17 LFIAT). También, se presentó en esta ley la positivización de un sandbox regulatorio, que establece en qué escenarios van a poderse crear desarrollos experimentales con limitaciones a la responsabilidad jurídica. Además, se presentan lineamientos de seguridad técnica que se harán en coordinación de la Agencia de Ciberseguridad del Estado que deberán cumplir los sistemas de inteligencia artificial.

Se presenta la inscripción en el Registro Nacional como un requisito indispensable para poder obtener las diferentes protecciones legales de la ley, mediante los parámetros técnicos necesarios, así como la obligación expresa del uso de trabajo de datos que sean exclusivamente de dominio abierto o de su propiedad, y que estos proyectos no tengan el propósito de fines comerciales ni para el público general (art. 16 LFIAT). La ley obliga tácitamente a los sistemas que mantengan fines comerciales que quieran beneficiarse de las salvaguardas que proporciona la Ley del mismo modo que lo hacen los mencionados en el artículo anterior. Mientras que, los que tengan fines comerciales tendrán la obligación de informar al usuario cuando las decisiones sean impulsadas por IA o cuando hubo una incidencia en la toma de decisiones. Se deberá explicar el proceso en la toma de decisiones, proporcionando una explicación comprensible y contar con mecanismos de impugnación mediante la posible apelación de una persona natural competente (art. 18 LFIAT).

Los principios éticos obligatorios no son definidos, sino únicamente mencionados y eso dificulta su comprensión y la determinación del techo normativo. Se especifican como parte de estos principios la equidad, entendiéndose como un trato justo y equilibrado, así como libre de sesgos; transparencia, como la claridad del funcionamiento y la explicación de decisiones de las personas físicas o jurídicas a quienes les afecta la ley; responsabilidad, se entiende como la rendición de cuentas, la evaluación de impactos y programaciones de auditorías; inclusión y no

¹¹⁸ Asamblea Legislativa de la República de El Salvador, Decreto nº 363, 18 de julio de 2025.

discriminación, ofrecería la participación de todas las partes interesadas; minimización de datos y consentimiento informado, se entiende como la autorización sobre el uso de los datos personales (art. 5 LFIAT).

Esto va de la mano de la protección de datos personales y en cierta medida, la ley de Ciberseguridad, en todo lo que le sea aplicable. El cumplimiento normativo se adherirá a la Ley para la Protección de Datos Personales y se implementarán medidas técnicas para la seguridad de datos. Se establecerán una serie de obligaciones que se centran en la no divulgación sin consentimiento de “algoritmos propietarios”, los cuales deberán dar protección a los *data sets* y a los sistemas de IA (art. 23 LFIAT). Del mismo modo que, se establecerán unas salvaguardas que se integrarán para aquellos que se inscriban en el registro y cumplan las normativas técnicas de la ANIA. Existirá una protección del uso de datos abiertos, espacios de experimentación segura sin responsabilidad en sandbox no comerciales, protección a desarrolladores por un uso indebido de terceros por posibles ataques de ciberseguridad y para cuestiones de exportación tecnológica para los usos que se den fuera de El Salvador no caigan sobre la responsabilidad de los desarrolladores.

No obstante, el hecho de que los sistemas de IA mantengan unas características opacas, complejas, con interconexión y autonomía, dificultan la regulación de la responsabilidad y establece una protección en base al derecho común, asignando responsabilidades según el rol específico en el ciclo de vida de la IA y estableciendo la responsabilidad de forma aislada en cada caso. Respecto a los desarrolladores, tendrán la obligación de diseñar sistemas éticos y técnicamente sólidos; los implementadores deberán garantizar la implementación correcta y segura; los proveedores deben mantener una infraestructura confiable; y los usuarios deberán usar la IA dentro su alcance previsto. Según el código civil salvadoreño, deberán hacerse cargo de la reparación de perjuicios tanto a nivel contractual como extracontractual.¹¹⁹

3. Peligros de la IA y afectación de Derechos Humanos en El Salvador

Las sofisticadas estrategias de marketing relacionadas con la recopilación de datos, las cuales también se aplican al ámbito político, han alcanzado niveles significativos.¹²⁰ Según afirman los investigadores Kosinski, Stillwell y Graepel, sería suficiente con 100 “me gusta” en publicaciones, comentarios o reels para elaborar un perfil ajustado del usuario del que se trata, revelando información confidencial, así como la ideología, orientación sexual, religión e incluso,

¹¹⁹ NUDO PROPIETARIO, “IA El Salvador: del marco legal al éxito empresarial”, 2025.

Disponible en: <https://www.youtube.com/watch?v=HkLpXUh9BCc>

¹²⁰ CASTELLANOS CLARAMUNT, J., “La Inteligencia Artificial en el contexto jurídico: El derecho como garante del desarrollo democrático”, *Diálogos Jurídicos*, 2024, p. 92.

extenderse hasta el nivel de inteligencia. Lo preocupante de este análisis de big data acaba recayendo sobre la precisión del algoritmo, pudiendo amenazar la privacidad de los usuarios y considerando que es el propio usuario el que decide, de manera consciente, a pesar de no tener suficiente consciencia de las posibles implicaciones de esos actos.¹²¹

Esta situación se ve trasladada también al espectro político. La influencia de diversos elementos tecnológicos como las redes sociales destacan por su amplio impacto sobre la capacidad de decisión de los usuarios reflejado en los procesos electorales, siendo únicamente necesarios extractos de las sesiones plenarias en forma de reels o posts, para poder hacer cambiar de opinión o fortalecerla mediante discursos que aboguen por la transmisión de emociones como el odio, el desencanto, la ira o la pasión, dejando en una posición subalterna la importancia de las ideas. Esta cuestión tampoco es novedosa, pues la presencia de emociones en los discursos políticos ha formado parte de los patrones de legitimación y manipulación de la política tradicional. El problema reside en la reducida resistencia que manifiestan los votantes y usuarios al no considerar la necesidad de contrastar la información que les viene dada como cierta. Este escenario se ve todavía más frágil con la llegada de la Inteligencia Artificial, dado que un individuo influyente puede aprovechar su posición para dirigir mensajes a segmentos específicos con una aparente objetividad y a través de una sugerencia que se adapte al sujeto. Esta aparente retórica discursiva puede ser parte de una manipulación política que desemboque en un cambio de gobierno o en la perpetuación en el poder de quien lo ostentase. Del mismo modo que sería indetectable que el ciudadano se diera cuenta del proceso, debido a la gran cantidad de datos proporcionados y haciendo útil la modulación sustancial de sus convicciones políticas.

La influencia a la que apuntamos determina un contexto que ejerce sobre el subconsciente del individuo y de la sociedad en el resultado de estrategias electorales enfocadas en la obtención de datos y en el microtargeting, refiriéndose a la segmentación de la audiencia en grupos reducidos según criterios demográficos, psicográficos, geográficos o, reforzados por una problemática común¹²², como el contexto de violencia que sufría El Salvador durante décadas. La Inteligencia Artificial a pesar de ser una importante herramienta en numerosas situaciones, también debe tratarse como una amenaza para las democracias en caso de que no se le dé el uso debido. Ante el nuevo antagonismo tecnológico que reordena nuestro mundo mediante la confrontación de China y Estados Unidos, se puede observar que para que exista un amplio

¹²¹ KOSINSKI, M., STILLWELL, D., GRAEPEL, T., <<Private traits and attributes are predictable from digital records of human behaviour>>, PNAS, vol. 110, n. 15, 2013.

¹²² CASTELLANOS CLARAMUNT, J., "La Inteligencia Artificial en el contexto jurídico: El derecho como garante del desarrollo democrático", *Diálogos Jurídicos*, 2024, pp.87 – 88.

protagonismo de las inteligencias artificiales no es necesario la existencia de un sistema democrático, sino que más bien puede resultar un obstáculo para su intención de desarrollo. Esto ha dado lugar a dos nuevas variantes del sistema económico, por una parte, el capitalismo de plataformas, donde las empresas desarrollan un mercado en concreto a partir de un algoritmo de oferta y demanda como principal activo físico. Y, por otra parte, el capitalismo de vigilancia, que se encarga de la extracción de datos de la actividad de los usuarios con el objeto de monetizarlos en campañas de publicidad y de marketing, del mismo modo que también puede llegar a ser usado para influir en votaciones electorales.¹²³

En un contexto como el de El Salvador que se rige por una figura populista como presidente y que concentra en gran medida el poder de toda la nación, muestra un escenario en el que acaba sucumbiendo las herramientas del Estado según sus intereses y a través de una dialéctica convincente que podría utilizar para su propio interés los beneficios de una inteligencia artificial. Si bien una regulación en esta materia es necesaria, no cabe olvidar que El Salvador se encuentra en un estado de excepción desde marzo de 2022 y que se extiende hasta la actualidad, suspendiendo las garantías constitucionales establecidas en los art. 12 inciso 2º, 13 inciso 2º y 24 de la Constitución de la República, para la continuidad del restablecimiento del orden, la seguridad ciudadana y el control territorial.¹²⁴

Con anterioridad ya se aprobó legislación similar en El Salvador en las que instituciones internacionales ya señalaron los peligros y posibles afectaciones de derechos humanos que podrían producirse bajo la aplicación de estas, tales como la libertad de expresión o la privacidad de aquellas personas que mostrasen una postura disconforme con el gobierno. Entre estas, se publicó el Decreto Legislativo nº144 relativo en la Ley para la Protección de Datos Personales, que configura la regulación de datos personales determinando los requisitos esenciales para realizar el tratamiento legítimo e informado sobre recolección, uso, procesamiento y almacenamiento de datos bajo el derecho a la intimidad y a la autodeterminación informativa de las personas naturales. Pero al mismo tiempo que la anterior ley podría atentar contra libertad de expresión, pudiendo provocar la eliminación de una noticia periodística sobre un funcionario público en caso de que considerase que las fuentes consultadas o con las cuales se obtuvieron la información se tratan de medios ilícitos y llegando a poder provocar una serie de procesos judiciales en contra de los periodistas.

¹²³ SEVILLA, J., *El pulgar inteligente y la democracia enferma*, Tirant lo Blanch, València, 2025, p. 110.

¹²⁴ Asamblea Legislativa de la República de El Salvador, Decreto nº 392, D.O. nº 162, Tomo nº 448, 2025.

Seguido a esta, el 12 de noviembre de 2024 se aprobó en la Asamblea Legislativa del Salvador la creación de la Agenda de Ciberseguridad del Estado (ACE), cuya se dirige por un director general nombrado por el presidente, y se fundamenta en la supervisión y regulación de los mecanismos de protección de datos. Esta ley de protección de datos ha manifestado el reconocimiento al “derecho al olvido”, que otorga extensas facultades a la nueva agencia estatal porque ordeno la eliminación de información sobre individuos en plataformas y páginas web. Además de atribuir esta ley la competencia para considerar si los datos son inadecuados, inexactos o excesivos, puede prever una excepción para la libertad de prensa, condicionada al cumplimiento de un principio de exactitud, lo cual puede dar lugar a interpretaciones restrictivas y producir una censura encubierta.

Según como apunta Human Rights Watch, estas leyes podrían llegar a facilitar la supresión de contenido crítico con el gobierno bajo el pretexto de proteger datos personales, provocando con esta vía, una carencia de transparencia y generando un entorno propicio para la censura. Es por este motivo por el cual se ha manifestado la ONU y otros organismos como la Relatoría Especial de la OEA para la Libertad de Expresión, advirtiendo que estos mecanismos de protección de datos tendrían que ser independientes y estar sujetos a garantías procesales sólidas.¹²⁵

La organización Artículo 19¹²⁶ mantuvo la misma línea sobre esta cuestión, pues el olvido no tendría que servir para eliminar información de interés público ni para restringir la libertad de prensa. Además que, la propia ley excluye su aplicación sobre los datos relacionados con seguridad y defensa, llevando a que se pueda difundir datos personales con justificaciones insustanciales, menguando la protección de la privacidad de los ciudadanos y de los críticos con el gobierno.¹²⁷ Aun así, estas leyes llegan a mantener cierta verosimilitud con las que aprobó Daniel Ortega en Nicaragua, las cuales han producido una gran restricción en cuanto a la libertad de prensa y de expresión, a la vez que se persiguen y se condenan en el país centroamericano, aquellos delitos considerados como cibercrímenes.¹²⁸

¹²⁵ HUMAN RIGHTS WATCH, “El Salvador: Nuevas leyes amenazan la libertad de expresión y la privacidad”, 2024. Disponible a: <https://www.hrw.org/es/news/2024/12/12/el-salvador-nuevas-leyes-amenazan-la-libertad-de-expresion-y-la-privacidad> (consultado el 10 de abril de 2025)

¹²⁶ ARTICLE 19, *The Global Principles on Protection of Freedom of Expression and Privacy*, (s.f.).

Disponible a: <https://www.article19.org/data/files/medialibrary/38657/expression-and-privacy-principles-1.pdf>

¹²⁷ HUMAN RIGHTS WATCH, “El Salvador: Nuevas leyes amenazan la libertad de expresión y la privacidad”, 2024.

Disponible a: <https://www.hrw.org/es/news/2024/12/12/el-salvador-nuevas-leyes-amenazan-la-libertad-de-expresion-y-la-privacidad> (Consultado el 10 de abril de 2025)

¹²⁸ MALDONADO, C.S., “Un informe de Human Rights Watch alerta de las amenazas de las <<leyes mordaza>> de Bukele en El Salvador”, *El País*, 2024.

Por otra parte, también se encuentra la Ley de Agentes Extranjeros que constituye una limitación indebida a los derechos fundamentales, así como el derecho de reunión, asociación y expresión. Esta ley, ya advertía en 2024 Angélica Cárcamo -presidenta de APES y directora de la Red Centroamericana de periodistas-, de la posible implementación en El Salvador del mismo modo que se hizo en la Nicaragua de Ortega.¹²⁹ Esta normativa implica que las personas físicas o jurídicas que reciben financiación extranjera para el desarrollo de las actividades en El Salvador¹³⁰, tendrán la obligación de inscribirse al Registro de Agentes Extranjeros (RAEX), con dependencia del Ministerio de Gobernación y Desarrollo Territorial, a cargo de un director general nombrado por la cartera del Estado. Este registro público tendrá carácter obligatorio sobre todos los sujetos obligados, pudiendo administrar por parte del gobierno los requisitos, así como la aprobación, denegación o no renovación de las solicitudes del registro.¹³¹ A pesar de esto, a la ley no se presenta ningún criterio firme que permita excluir a las organizaciones de este registro, pudiendo afectar a la inclusión o exclusión de aquellas organizaciones que lleguen a incomodar al gobierno, tratándose de activistas de la sociedad civil, defensoras y defensores de derechos humanos o periodistas. Este registro, según las afirmaciones de la Asamblea Legislativa, pretende canalizar los recursos financieros o activos recibidos de sujetos, entidades o instituciones que estén o no supervisadas por la Superintendencia del Sistema Financiero. Siguiendo el objetivo de poder llegar a llevar correctamente los libros de contabilidad, a la vez que se cumple con la Ley Contra el Lavado de Dinero y de Activos. Aun así, se impondrán graves multas y sanciones sobre aquellas organizaciones o agentes afectados que incumplan las obligaciones y plazos establecidos para presentar la información requerida y enmendar las prevenciones.¹³²

Por otro lado, Nayib Bukele justifica la aprobación de esta normativa mediante una publicación en la red social X, en la cual sostiene que la carencia de control efectivo sobre las organizaciones no gubernamentales, junto con la dinámica política tradicional de la FMLN y el contexto de las reformas agrarias en Latinoamérica, ha propiciado el establecimiento de contratos fraudulentos entre actores políticos y empresarios. Estos entramados, según Nayib Bukele, han contado con la colaboración de entidades, así como ONGs, incitando en las comunidades a

Disponible a: <https://elpais.com/america/2024-12-12/un-informe-de-human-rights-watch-alerta-de-las-amenazas-de-las-leyes-mordaza-de-bukele-en-el-salvador.html> (consultado el 12 de abril de 2025)

¹²⁹ CÁRCAMO, A., *¿Hacia dónde va la libertad de expresión en El Salvador tras reelección de Bukele?*, Universitat de València, Facultad de Ciencias Sociales, 25 de abril de 2024.

¹³⁰ CNN, "¿Qué busca Bukele con la ley de agentes extranjeros?", 2025.

Disponible en: https://www.youtube.com/watch?v=f_wt9fnyjeu (Consultado el 26 de mayo de 2025)

¹³¹ ASAMBLEA LEGISLATIVA DE EL SALVADOR, "Ley de Agentes Extranjeros busca transparentar uso de fondos provenientes del exterior", 2025.

Disponible en: <https://www.asamblea.gob.sv/node/13572> (Consultado el 26 de mayo de 2025)

¹³² Ibid

protestar contra una propiedad residencial privada por la pérdida de las tierras. Mientras que, el caso que menciona estuvo vinculado a procesos de transferencia irregular de tierras salvadoreñas a favor de Luis Palomo, de Óscar Ortiz (exalcalde de Santa Tecla por la FMLN) y, posteriormente, de la administración municipal encabezada por Roberto de Aubuisson, del partido ARENA.¹³³ A pesar de que se presenta esta ley con la intención de salvaguardar la soberanía nacional y favorecer una mayor transparencia, las bases de esta ley están siendo dirigidas para crear una herramienta de control porque aquellas organizaciones que no estén al registro, tengan el temor y obedezcan los lineamientos del gobierno, para mantenerse fuera de este y no tener que pagar el impuesto del 30% que establece esta norma.¹³⁴

Asimismo, las diferentes nuevas legislaciones aplicadas actualmente en El Salvador, entre ellas las recientes reformas constitucionales, harían posible que esta excepcionalidad permanezca todavía más normalizada. Tratando de hacer creer a los ciudadanos y analistas internacionales, que este régimen de excepción supone una indispensable necesidad para asegurar la actual seguridad y la conseguida reducción de criminalidad en el territorio. El modelo que plantea Bukele, supone una legitimación de una receta autoritaria, la cual se conforma a partir de una mala calidad de seguridad que se basa al romper con los pilares fundamentales del Estado de derecho. Pero que resalta la situación salvadoreña sobre la cuestión que se plantea ante la experiencia estructural de una inseguridad extrema, que dispone el apoyo popular a todo aquel gobernante que ofrezca un alivio a la situación tan extrema de vida cotidiana. Esto plantea la actual disyuntiva en la cual, la cuestión criminal toma papel a la hora de determinar el modelo político, a pesar de que su precio sea la anulación de garantías básicas o la alteración de los parámetros básicos de una democracia en la que se sitúa la división de poderes y el sistema de contrapesos.¹³⁵

Conclusión

El desarrollo de las nuevas tecnologías y de la inteligencia artificial en los espacios de gobernanza contemporánea está marcado por los esfuerzos que requieren su armonización y regulación. En este marco se genera una intensa discusión legislativa y normativa que enfrenta

133 BUKELE, N. (nayibbukele), *“Un ejemplo de cómo funcionan las ONGs, la política tradicional y las reformas agrarias en Latinoamérica: ¿Cómo se pierden 384 manzanas de tierra? ¿A la fuerza? No. ¿Con balas? No. Empiezo...”*, X, 16 de mayo de 2025.

Disponible a: <https://x.com/nayibbukele/status/1923207408949625009?t=h6wkiv2b74zkcxl9pxbmvw&s=19>
(Consultado el 26 de mayo de 2025)

134 CNN, *“¿Qué busca Bukele con la ley de agentes extranjeros?”*, 2025

135 ALFIE, J., *“La falsa eficacia del modelo Bukele”*, *Revista Anfibia & Universidad Nacional de San Martín*, 2024.
Disponible a: <https://www.revistaanfibia.com/la-falsa-eficacia-del-modelo-bukele/>

la fragmentación regulatoria existente, una realidad que también alcanza a la República de El Salvador.

La aprobación de la ley de fomento de la inteligencia artificial y las tecnologías de El Salvador supone ciertas incógnitas de cómo será su desarrollo y aplicación en el territorio. Según la plataforma de transparencia del Estado, dicha normativa busca mejorar los servicios públicos e impulsar la economía nacional. Asimismo, establece su ámbito de aplicación en personas naturales y jurídicas involucradas en el desarrollo, investigación y aplicación de estas tecnologías. Entre las novedades introducidas se encuentra la creación de nuevas instituciones, como la Agencia Nacional de Inteligencia (ANIA), y la incorporación de principios éticos relacionados con la equidad, la transparencia, la responsabilidad y la inclusión.

No obstante, el texto legal no profundiza en los riesgos vinculados al uso de estas tecnologías ni delimita con claridad los espacios en los que podrán ser implementadas. Esta omisión genera preocupación respecto a la posibilidad de que la legislación sea utilizada con fines que favorezcan al gobierno en detrimento de los derechos humanos. En este sentido, resulta fundamental analizar de qué manera la normativa puede contribuir al fortalecimiento de la gobernanza democrática y al desarrollo de la sociedad civil, al mismo tiempo que se reconoce el riesgo de que sea instrumentalizada para la consolidación de prácticas autoritarias o para la limitación de libertades fundamentales.

En conclusión, esta legislación representa un paso inicial y relevante hacia la incorporación de la inteligencia artificial en la estructura institucional de El Salvador. Sin embargo, al tratarse de una norma de primera generación, presenta carencias importantes en cuanto a la identificación de riesgos y la protección efectiva de derechos. El futuro de esta iniciativa dependerá no solo de su correcta implementación, sino también de la capacidad de las instituciones salvadoreñas para garantizar que la innovación tecnológica se convierta en una herramienta para el bienestar colectivo, y no en un instrumento de retroceso democrático o vulneración de derechos humanos.

Bibliografía

ALFIE, J., “La falsa eficacia del modelo Bukele”, *Revista Anfibia & Universidad Nacional de San Martín*, 2024.

Disponible a: <https://www.revistaanfibia.com/la-falsa-eficacia-del-modelo-bukele/>

ARTICLE 19, *The Global Principles on Protection of Freedom of Expression and Privacy*, (s.f.).

Disponible a: <https://www.article19.org/data/files/medialibrary/38657/expression-and-privacy-principles-1.pdf>

Asamblea Legislativa de la República de El Salvador, *Decreto nº 363*, 18 de julio de 2025.

-: *Decreto nº 392*, D.O. nº 162, Tomo nº 448, 2025.

-: *Ley de fomento a inteligencia artificial y tecnologías*, Decreto nº234, 3 de marzo de 2025.

-: “Ley de Inteligencia Artificial ayudará a mejorar los servicios públicos e impulsar la economía”, 2025.

Disponible en: <https://www.asamblea.gob.sv/node/13492#:~:text=La> (Consultado el 22 de septiembre de 2025)

-: “Ley de Agentes Extranjeros busca transparentar uso de fondos provenientes del exterior”, 2025.

Disponible en: <https://www.asamblea.gob.sv/node/13572> (Consultado el 26 de mayo de 2025)

BUKELE, N. (nayibbukele), “Un ejemplo de cómo funcionan las ONGs, la política tradicional y las reformas agrarias en Latinoamérica: ¿Cómo se pierden 384 manzanas de tierra? ¿A la fuerza? No. ¿Con balas? No. Empiezo...”, X, 16 de mayo de 2025.

Disponible a: <https://x.com/nayibbukele/status/1923207408949625009?t=h6wkiv2b74zkcxl9pxbmvw&s=19>

(Consultado el 26 de mayo de 2025)

CÁRCAMO, A., *¿Hacia dónde va la libertad de expresión en El Salvador tras reelección de Bukele?*, Universitat de València, Facultad de Ciencias Sociales, 25 de abril de 2024.

CASTELLANOS CLARAMUNT, J., “La Inteligencia Artificial en el contexto jurídico: El derecho como garante del desarrollo democrático”, *Diálogos Jurídicos*, 2024, pp. 85 – 101.

CENTRO NACIONAL DE INTELIGENCIA ARTIFICIAL (CENIA), *Índice Latinoamericano de Inteligencia Artificial*, 2024.

CNN, “¿Qué busca Bukele con la ley de agentes extranjeros?”, 2025.

Disponible en: https://www.youtube.com/watch?v=f_wt9fnyjeu (Consultado el 26 de mayo de 2025)

Congreso de la República de Perú, *Ley que promueve el uso de la inteligencia artificial en favor del desarrollo económico y social del país*, Ley nº31814, 2023.

Disponible en: <https://cdn.www.gob.pe/uploads/document/file/5038703/ley-que-promueve-el-uso-de-la-inteligencia-artificial-en-fav-ley-n-31814.pdf?v=1692895308>

CORVARÁN, J. G. & CARRO, M.V., “El Reglamento de inteligencia artificial desde fuera de la Unión Europea: impulsos reguladores desde otras partes del mundo y una visión desde Iberoamérica”, en *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Editorial Aranzadi, 2024, pp. 85 – 109.

GARCÍA CABRERA, B.A., “Inteligencia Artificial y economía”, *Revista Ciencias Económicas*, 1 (1), San Salvador, El Salvador, 2023, pp. 31-36.

HUMAN RIGHTS WATCH, “El Salvador: Nuevas leyes amenazan la libertad de expresión y la privacidad”, 2024.

Disponible a: <https://www.hrw.org/es/news/2024/12/12/el-salvador-nuevas-leyes-amenazan-la-libertad-de-expresion-y-la-privacidad> (consultado el 10 de abril de 2025)

LÓPEZ, G., “Diputada Ortiz pide un debate amplio sobre la propuesta de ley de fomento de la inteligencia artificial “en términos éticos y económicos”, *LaNoticiaSV*, 2025.

Disponible en: <https://lanoticiasv.com/diputada-ortiz-pide-un-debate-amplio-sobre-la-propuesta-de-ley-de-fomento-de-la-inteligencia-artificial-en-terminos-eticos-y-economicos/> (Consultado el 24 de septiembre de 2025)

MALDONADO, C.S., “Un informe de Human Rights Watch alerta de las amenazas de las <<leyes mordaza>> de Bukele en El Salvador”, *El País*, 2024.

Disponible a: <https://elpais.com/america/2024-12-12/un-informe-de-human-rights-watch-alerta-de-las-amenazas-de-las-leyes-mordaza-de-bukele-en-el-salvador.html> (consultado el 12 de abril de 2025)

MANTELERO, A., “El Reglamento de inteligencia artificial: la respuesta del legislador europeo a los retos de la inteligencia artificial”, en *Tratado sobre el Reglamento de Inteligencia Artificial de la Unión Europea*, Editorial Aranzadi, 2024, pp. 53 – 66.

NUDO PROPIETARIO, “IA El Salvador: del marco legal al éxito empresarial”, 2025.

Disponible en: <https://www.youtube.com/watch?v=HkLpXUh9BCc>

SEVILLA, J., *El pulgar inteligente y la democracia enferma*, Tirant lo Blanch, València, 2025.

KOSINSKI, M., STILLWELL, D., GRAEPEL, T., <<Private traits and attributes are predictable from digital records of human behaviour>>, *PNAS*, vol. 110, n. 15, 2013.

[ACCEDE A LA GRABACIÓN](#)

Resumen

Este trabajo analiza el impacto de la opacidad algorítmica sobre los derechos humanos en el contexto normativo de la Unión Europea (UE) y Estados Unidos (EE.UU.). A partir de un enfoque jurídico-crítico, se argumenta que la opacidad no es únicamente un fenómeno técnico, sino una condición estructural que impide la rendición de cuentas, obstaculiza la trazabilidad de las decisiones automatizadas y vulnera principios fundamentales como la igualdad, la no discriminación, la privacidad y el debido proceso. Se examina cómo la UE, mediante el reglamento 2024/1689, ha adoptado un enfoque preventivo basado en obligaciones legales de transparencia, inteligibilidad y evaluación de impacto en derechos fundamentales. Por el contrario, se demuestra que el marco estadounidense permanece fragmentado, dependiente de regulaciones sectoriales o estatales, sin una arquitectura jurídica coherente a nivel federal. A través de casos emblemáticos (Syri, lborderctrl, Compas, Saferent), se ilustra la materialización práctica de la opacidad y sus efectos desproporcionados sobre grupos vulnerables. Finalmente, se destaca la asimetría tecnológica global, subrayando que la mayoría de los sistemas de IA desplegados en la UE son desarrollados en EE.UU., lo que plantea desafíos de aplicación extraterritorial del derecho europeo y limita su eficacia real. Frente a ello, se propone una serie de principios normativos comunes, orientados a consolidar una gobernanza algorítmica garantista basada en la trazabilidad, la inteligibilidad y el control democrático, como condiciones indispensables para garantizar una protección estructural de los derechos humanos en entornos cada vez más automatizados.

1. Marco conceptual: opacidad algorítmica, transparencia, trazabilidad e inteligibilidad

En el ámbito de la Inteligencia Artificial (IA), el concepto de opacidad algorítmica no debe reducirse a la inaccesibilidad del código fuente ni a la mera complejidad técnica de los modelos, sino que ha de entenderse como un fenómeno estructural, multifactorial y transversal, que condiciona de manera decisiva la comprensión, supervisión y control de los procesos de decisión automatizada. Esta opacidad no surge únicamente de las características técnicas de los sistemas, sino también de factores jurídicos, económicos y organizativos, como el secreto comercial, la asimetría de información entre proveedores y usuarios, o la ausencia de obligaciones regulatorias claras.

Desde una perspectiva analítica, la opacidad algorítmica puede descomponerse en al menos cuatro dimensiones interrelacionadas. En primer lugar, la falta de transparencia, entendida como la imposibilidad de conocer los datos de entrada, los criterios de funcionamiento o los resultados del sistema. En segundo lugar, la ausencia de trazabilidad, que impide reconstruir la cadena decisional seguida por el algoritmo, dificultando tanto la auditoría interna como el control externo. En tercer lugar, la ininteligibilidad, que se produce cuando los resultados o razonamientos del sistema no pueden ser comprendidos ni por los usuarios afectados ni por los operadores humanos que interactúan con la herramienta. Finalmente, se encuentra la limitación en la auditabilidad, es decir, la dificultad para someter el sistema a escrutinio técnico, jurídico o institucional independiente, ya sea por falta de acceso a la documentación, por restricciones contractuales o por la complejidad inherente del modelo.

En este contexto, los denominados sistemas de “caja negra” (*black box systems*) representan una forma paradigmática de opacidad: en ellos, solo son accesibles las entradas y salidas, mientras que los procesos intermedios permanecen opacos, incluso en ocasiones para los propios desarrolladores. Sin embargo, conviene precisar que no toda opacidad equivale a “caja negra”. Un sistema puede resultar incomprensible o inaccesible para los usuarios finales, aun cuando sus diseñadores sí dispongan de mecanismos internos de explicación. Asimismo, la opacidad debe diferenciarse del sesgo algorítmico: aunque ambos fenómenos se encuentran relacionados, la opacidad no implica necesariamente parcialidad o discriminación, pero sí puede encubrir, amplificar o impedir la detección de sesgos existentes, con efectos directos sobre la igualdad y la no discriminación.

Resulta igualmente relevante distinguir entre explicabilidad post hoc (mecanismos para interpretar los resultados de un modelo ya entrenado, como técnicas de *IA explicable*) e interpretabilidad ex ante (diseño de modelos intrínsecamente comprensibles). A ello debe añadirse una tercera capa: la transparencia procedimental y documental, que incluye registros de datos, documentación técnica y justificación de las decisiones de diseño. Estas tres

dimensiones no son equivalentes, y una regulación eficaz debería exigir las de forma complementaria para garantizar un control real.

Por otra parte, es necesario enfatizar que los algoritmos no deciden por sí mismos, sino que ejecutan instrucciones humanas enmarcadas en objetivos, parámetros y conjuntos de datos definidos por actores concretos. El problema de la opacidad no radica en una supuesta autonomía tecnológica de la inteligencia artificial, sino en la ausencia de responsabilidad institucional y en la falta de mecanismos jurídicos que aseguren la rendición de cuentas. Precisamente por ello, la lucha contra la opacidad algorítmica (a través de exigencias normativas de transparencia, trazabilidad, inteligibilidad y auditabilidad) constituye un eje esencial para la defensa de los derechos humanos en la era digital.

2. Derechos humanos afectados por la opacidad algorítmica

La expansión de los sistemas algorítmicos en sectores clave como la justicia, la seguridad, el empleo, la sanidad o la gestión pública ha generado una preocupación creciente sobre los derechos fundamentales comprometidos cuando dichas tecnologías operan bajo condiciones de opacidad. Este fenómeno no afecta de manera uniforme a todos los derechos, sino que incide en núcleos esenciales del Estado de derecho y de la protección internacional de los derechos humanos.

En primer lugar, la igualdad y la no discriminación (art. 21 Carta de Derechos Fundamentales de la UE, CDFUE en adelante) se ven comprometidas cuando los algoritmos perpetúan sesgos ocultos (raciales, de género, socioeconómicos o etarios) que resultan invisibles para los afectados y difícilmente impugnables. La opacidad impide identificar cómo se han construido los conjuntos de entrenamiento y qué criterios han condicionado los resultados, lo que genera exclusión estructural. Estos efectos contravienen también el principio de equidad en el tratamiento de datos (art. 5 RGPD) [1].

En segundo lugar, se ve afectado el derecho a un juicio justo y al debido proceso (art. 47 CDFUE y art. 6 del Convenio Europeo de Derechos Humanos, CEDH en adelante), particularmente en contextos judiciales, policiales o migratorios. Cuando las decisiones automatizadas influyen en detenciones preventivas, concesión de libertades o resoluciones administrativas, la falta de explicaciones vulnera el derecho a ser oído y a contar con recursos efectivos. La Directiva (UE) 2016/680 sobre tratamiento de datos en el ámbito policial y penal refuerza esta exigencia, al establecer límites claros al uso de sistemas automatizados en contextos de seguridad pública [2].

Asimismo, el derecho a la privacidad y a la protección de datos personales (arts. 7 y 8 CDFUE, art. 8 CEDH) se ve erosionado en escenarios de opacidad algorítmica. El RGPD, en sus artículos 12 a 15, reconoce el derecho de los interesados a recibir información clara sobre el tratamiento

de sus datos, y en el artículo 22 limita las decisiones basadas exclusivamente en procesos automatizados con efectos jurídicos relevantes. Sin embargo, la opacidad técnica y contractual hace que estos derechos resulten de difícil ejercicio en la práctica [3].

Otro derecho afectado es el derecho a la buena administración y al acceso a la información (art. 41 y 42 CDFUE), especialmente en el ámbito público, donde las administraciones utilizan algoritmos para gestionar prestaciones sociales, asignar recursos o ejercer funciones de control sin notificación previa ni explicación suficiente. Esta práctica debilita la transparencia administrativa y la confianza en las instituciones democráticas [4].

Finalmente, en contextos democráticos, la falta de inteligibilidad estructural amenaza el derecho a la participación en los asuntos públicos (art. 39 CDFUE y art. 25 del Pacto Internacional de Derechos Civiles y Políticos). Si la ciudadanía no puede comprender cómo operan los sistemas que condicionan el acceso a derechos y servicios, la calidad del proceso democrático se ve gravemente erosionada [5].

Frente a estas amenazas, el desafío no consiste en restringir el uso de la inteligencia artificial, sino en dotar al ecosistema digital de garantías técnicas, normativas e institucionales que aseguren trazabilidad, auditabilidad y rendición de cuentas, permitiendo que los derechos fundamentales puedan ejercerse en un entorno crecientemente automatizado [6].

3. La respuesta normativa de la Unión Europea frente a la opacidad algorítmica

La Unión Europea ha situado la transparencia algorítmica y la protección de los derechos fundamentales en el centro de su estrategia reguladora en materia de inteligencia artificial. El Reglamento (UE) 2024/1689 sobre Inteligencia Artificial (RIA, en lo que sigue) constituye el eje normativo de esta apuesta, consolidando un enfoque basado en la gestión del riesgo, en el que las obligaciones jurídicas se establecen de manera proporcional al impacto potencial de los sistemas en juego.

Para los sistemas de IA de alto riesgo, el RIA impone un conjunto de obligaciones específicas que buscan hacer frente a la opacidad algorítmica. En particular, el artículo 13 exige garantizar que los sistemas sean “suficientemente transparentes” para que los usuarios puedan interpretar sus resultados y utilizarlos adecuadamente. El artículo 14 establece la necesidad de supervisión humana efectiva, de forma que se evite una dependencia ciega de decisiones automatizadas. Por su parte, el artículo 16, junto con el anexo IV, obliga a los proveedores a elaborar una documentación técnica detallada sobre el diseño, el funcionamiento y la lógica del sistema, incluyendo la descripción de los datos empleados, las métricas utilizadas, sus limitaciones y los posibles impactos adversos sobre los derechos fundamentales [7].

Además, el RIA incorpora una herramienta de gran relevancia: la Evaluación de Impacto relativa a los Derechos Fundamentales, prevista en el artículo 27, aplicable a los sistemas de alto riesgo utilizados por autoridades y entidades del sector público. Esta evaluación debe identificar y mitigar los riesgos que dichos sistemas puedan suponer para la igualdad, la privacidad, la no discriminación o el derecho a un juicio justo antes de su despliegue efectivo. Su concepción está inspirada en la lógica del artículo 35 del Reglamento General de Protección de Datos (RGPD, en lo que sigue), pero con un alcance más amplio que incluye derechos no estrictamente vinculados a la privacidad, reforzando así el marco de garantías [6].

El RIA se articula, de este modo, con el resto del ordenamiento jurídico europeo: la Carta FDUE (arts. 7, 8, 21, 41 y 47) y el RGPD (arts. 12 a 15 y 22) proporcionan la base sustantiva sobre la que se erigen estas obligaciones de inteligibilidad, transparencia y control humano [7]. La combinación de estos instrumentos refleja una estrategia jurídica preventiva frente a la opacidad algorítmica, alineada con el principio de precaución que caracteriza al modelo regulatorio europeo.

No obstante, la eficacia de este marco normativo plantea varios desafíos. En primer lugar, persiste una indeterminación operativa en conceptos clave como “inteligibilidad” o “explicabilidad suficiente”, cuya interpretación práctica dependerá de directrices futuras y de la jurisprudencia. En segundo lugar, la implementación real de estas obligaciones depende de la capacidad técnica, jurídica y presupuestaria de las autoridades supervisoras en los Estados miembros, que deberán vigilar, auditar y sancionar el incumplimiento en un contexto de asimetría tecnológica respecto a los grandes desarrolladores privados [8].

En consecuencia, aunque el RIA constituye el modelo normativo más ambicioso hasta la fecha en materia de transparencia algorítmica, su impacto real dependerá de la solidez institucional de los mecanismos de control y de la disposición de los Estados miembros para traducir sus disposiciones en garantías efectivas y exigibles para los ciudadanos.

4. Ejemplos reales de opacidad algorítmica en la Unión Europea

A pesar de que regulación previa al RIA como la Digital Services Act (DSA) (por ejemplo, arts 27 y 28) la Digital Markets Act (DMA), la Data Governance Act (DGA, 2022) o la Data Act (2023/2024) incidían, de manera explícita en la importancia de la transparencia algorítmica, diversos casos han puesto de manifiesto que la opacidad de los sistemas de IA sigue comprometiendo derechos fundamentales en la práctica, haciendo necesaria una regulación amplia y completa como la desarrollada por la UE en el RIA. Algunos ejemplos son los siguientes:

SyRI (System Risk Indicator). Este sistema de evaluación de riesgos fue implementado en los Países Bajos para detectar fraude en prestaciones sociales; combinaba grandes volúmenes de datos personales y aplicaba modelos predictivos sin ofrecer a los ciudadanos ninguna

explicación sobre las decisiones adoptadas. En 2020, el Tribunal de Distrito de La Haya declaró que la legislación que amparaba SyRI vulneraba, entre otros, el artículo 8 del CEDH, al establecer una forma de vigilancia masiva carente de garantías de transparencia, proporcionalidad y control independiente [9]. La sentencia supuso un hito al subrayar que la mera eficiencia administrativa no puede justificar la utilización de sistemas opacos que afectan a derechos fundamentales [10].

iBorderCtrl. Este proyecto, Financiado por la Comisión Europea, buscaba aplicar un supuesto “detector de mentiras” basado en microexpresiones faciales y aprendizaje automático para evaluar la veracidad de los viajeros en pasos fronterizos. Desde su inicio, fue criticado por la comunidad científica y por organizaciones de derechos humanos debido a su falta de base empírica y a la opacidad de su funcionamiento. En 2019 se interpuso el recurso T-158/19 ante el Tribunal General de la Unión Europea, solicitando el acceso a la documentación técnica del sistema, que resolvió en 2021 permitiendo el acceso parcial a dichos documentos, aunque manteniendo restricciones por motivos de secreto comercial [11]. Este litigio reveló el conflicto entre el derecho a la transparencia y los intereses económicos de los desarrolladores, evidenciando las dificultades para someter proyectos tecnológicos de alto impacto a un control democrático efectivo [12].

Estos ejemplos muestran que, incluso bajo un marco normativo exigente como el europeo, la implementación práctica de sistemas de IA puede seguir generando zonas de opacidad institucionalizadas, donde los derechos a la privacidad, a la igualdad y al debido proceso quedan erosionados. La experiencia de SyRI e iBorderCtrl confirma que el reto no consiste únicamente en legislar la transparencia, sino en dotar a los tribunales, supervisores y a la ciudadanía de herramientas efectivas para garantizarla frente a la complejidad técnica y las presiones corporativas.

5. La fragmentación normativa en Estados Unidos ante la opacidad algorítmica

A diferencia del enfoque regulador unificado de la Unión Europea, el marco normativo de Estados Unidos frente a los sistemas de inteligencia artificial se caracteriza por su fragmentación, dispersión y fuerte influencia del sector privado. No existe, a nivel federal, una ley general que imponga obligaciones de transparencia, inteligibilidad o trazabilidad en el uso de algoritmos, ni un organismo central con autoridad para supervisar transversalmente su impacto sobre los derechos fundamentales [13].

Las regulaciones aplicables dependen del sector (financiero, sanitario, laboral, etc.) o del estado federado, lo que genera vacíos normativos y desigualdades en la protección. Un ejemplo pionero es la *Biometric Information Privacy Act* (BIPA) de Illinois (2008), que establece obligaciones de consentimiento informado y derecho a litigar contra el uso indebido de datos biométricos. Aunque inicialmente fue considerada una norma garantista, reformas recientes

han limitado el alcance de las demandas colectivas y reducido su eficacia práctica [14]. Otro ejemplo es la *Local Law 144* de Nueva York (2021), que exige auditorías de sesgo y notificación previa en el uso de sistemas automatizados de contratación. Sin embargo, su implementación ha sido lenta y las sanciones poco disuasorias, en parte debido a la presión de grupos empresariales [15].

A nivel federal, la *Federal Trade Commission* (FTC) ha publicado guías sobre la necesidad de transparencia, explicabilidad y equidad en el uso de sistemas de decisión automatizada. Estas orientaciones, sin embargo, carecen de fuerza vinculante y solo permiten sanciones en casos de prácticas desleales o engañosas bajo la legislación de consumo, sin configurar un marco integral para la gobernanza algorítmica [16].

Uno de los ejemplos paradigmáticos de opacidad es el caso del sistema COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), utilizado en varios estados para evaluar el riesgo de reincidencia penal. Aunque no constituía una “caja negra” en sentido estricto, la empresa propietaria, se negó a divulgar su funcionamiento alegando secreto comercial. Esto impidió a los acusados comprender los criterios que sustentaban su calificación de riesgo, afectando al derecho al debido proceso protegido por la Constitución de EE. UU. [17].

En conjunto, el modelo estadounidense no impone requisitos estructurales de inteligibilidad algorítmica y deja la transparencia en manos de la autorregulación empresarial o de iniciativas fragmentadas. Este diseño refuerza una lógica de mercado en la que los derechos fundamentales quedan supeditados a la discrecionalidad de los proveedores tecnológicos y a litigios aislados, más que a un control jurídico estructural.

6. Ejemplos reales de opacidad algorítmica en Estados Unidos

La opacidad algorítmica en Estados Unidos ha generado múltiples controversias con impacto en justicia penal, vivienda y empleo. Además del mencionado COMPAS, existen muchos casos representativos de falta de transparencia, trazabilidad e inteligibilidad, y de conflictos con normas antidiscriminatorias; algunos de ellos:

iTutorGroup. La Comisión para la Igualdad de Oportunidades en el Empleo, EEOC, resolvió en 2023 el que consideró su primer caso de IA discriminatoria, al alegar que el software de contratación desarrollado por la empresa mencionada rechazaba automáticamente a candidatas de más de 55 años y candidatos de más de 60 años. El acuerdo incluyó el pago de una compensación y ciertas acciones correctivas y preventivas, y evidenció cómo una lógica algorítmica no transparente puede traducirse en discriminación por edad sin vías de impugnación efectivas [20]. Este tipo de prácticas ha motivado respuestas regulatorias locales como la *Local Law 144 de NYC (AEDT)*, que exige auditorías de sesgo y avisos a candidatos [15].

SafeRent. En Massachusetts, diversas acciones colectivas denunciaron que la herramienta algorítmica *SafeRent Score*, utilizada para la evaluación automatizada de inquilinos, penalizaba de forma opaca a solicitantes con subvenciones de vivienda, generando efectos desproporcionados sobre personas negras e hispanas. En 2024 se alcanzó un acuerdo que impone la modificación del sistema y compensaciones económicas, reconociendo los riesgos de discriminación derivados de modelos no auditables en decisiones de acceso a vivienda [21–22]. El caso ilustra la complejidad de trazar y justificar puntuaciones automatizadas con consecuencias materiales.

En el contexto sanitario, diversos estudios han mostrado la existencia de diversos sesgos en aplicaciones comerciales. Por ejemplo [21] demostró la existencia de un sesgo racial en un algoritmo comercial de estratificación de pacientes ampliamente usado en EE.UU.: a igual riesgo asignado, los pacientes negros recibían menos recursos que pacientes blancos, porque la variable objetivo (el coste sanitario) ocultaba necesidades clínicas. Este es un ejemplo paradigmático de cómo la elección de variables y un diseño no transparentes inducen a resultados inequitativos.

En un contexto de privacidad y vigilancia, en 2022, la Unión Americana de Libertades Civiles (ACLU) alcanzó un acuerdo bajo la BIPA de Illinois que limitó la comercialización privada de la base de datos de *Clearview AI* y creó mecanismos de exclusión para residentes, subrayando la tensión entre secreto comercial y control público sobre sistemas altamente intrusivos y opacos [22]. En el mismo año, el Departamento de Justicia alcanzó un acuerdo con Meta por la segmentación algorítmica de anuncios de vivienda, obligando a modificar herramientas de audiencia por su potencial discriminatorio [23].

En conjunto, estos casos muestran que, sin una ley federal integral, la opacidad se dirime mediante litigios aislados, normas estatales específicas (por ejemplo, BIPA) y ordenanzas locales (por ejemplo, AEDT de Nueva York), lo que conduce a protecciones desiguales y a dificultades estructurales para auditar y explicar decisiones automatizadas con efectos reales sobre derechos fundamentales.

7. Comparación crítica entre los modelos de la UE y EE. UU.

La comparación entre la Unión Europea y Estados Unidos en materia de opacidad algorítmica revela dos arquitecturas regulatorias y culturas jurídicas con lógicas difícilmente conciliables en el corto plazo. La UE ha optado por un marco horizontal, ex ante y basado en riesgo, que convierte en obligaciones jurídicas exigibles la transparencia, la trazabilidad, la inteligibilidad y la supervisión humana en sistemas de alto riesgo (documentación técnica, registros, control humano, información significativa al usuario), bajo el paraguas del Reglamento (UE) 2024/1689 y su integración con la Carta DFUE y el RGPD [7]. EE. UU. mantiene, por su parte, un mosaico descentralizado y sectorial, donde la transparencia e inteligibilidad dependen de normas

estatales o municipales y de orientaciones federales no vinculantes (por ejemplo, guías de la FTC), lo que genera protecciones desiguales y remedios ex post principalmente a través de litigios y de la acción de agencias con competencias limitadas [16]. Esta divergencia puede desglosarse en varios ejes comparativos que ilustran las diferencias estructurales entre ambos modelos:

1.- Filosofía regulatoria y momento de control. La UE prioriza el control ex ante (conformidad previa a comercialización, obligaciones de diseño y documentación, evaluación de impacto en derechos en el sector público), buscando prevenir la opacidad antes del despliegue. EE.UU. confía en el control ex post (cumplimiento por orientación, aplicación *-enforcement-* caso a caso, responsabilidad civil y de consumo), de modo que la opacidad emerge como un riesgo aceptado hasta que un caso concreto demuestra su ilicitud o impacto.

2.- Estándar mínimo común. El modelo europeo establece un suelo normativo uniforme para toda la UE (obligaciones transversales y específicas por riesgo). El estadounidense carece de un marco mínimo común *-baseline-* federal, lo determinan leyes estatales o municipales y estándares voluntarios, con asimetrías territoriales y de sector, lo que afecta a la previsibilidad regulatoria y a la tutela homogénea de derechos.

3.- Transparencia e inteligibilidad. En la UE, la transparencia es deber legal con contenidos mínimos (información significativa, registros, documentación, límites a decisiones exclusivamente automatizadas), y la inteligibilidad funcional se convierte en requisito operativo para el uso responsable de salidas del sistema. En EE. UU., la transparencia se articula mediante políticas internas, avisos al consumidor y, en su caso, auditorías locales/estatales. La inteligibilidad depende en gran parte de la autorregulación y de acuerdos con autoridades, sin un derecho subjetivo general a la explicación.

4.- Protección del secreto comercial vs. escrutinio público. Ambos modelos reconocen la protección del secreto comercial, pero en la UE el acceso confidencial a autoridades y auditores está normativamente previsto. En EE. UU., el secreto comercial suele prevalecer hasta que el litigio o una investigación específica imponen aperturas parciales. De este modo, la UE dispone de una capacidad de supervisión más estructural y anticipatoria, mientras que el modelo estadounidense depende en mayor medida de la acción judicial ex post.

5.- Gobernanza e instituciones. La UE construye una infraestructura institucional (autoridades nacionales competentes, vigilancia de mercado, coordinación europea) para supervisar y sancionar, su reto es capacitar técnicamente a esos órganos. En EE. UU., la gobernanza reposa en agencias dispersas (FTC, EEOC, CFPB, HUD, etc.), con mandatos sectoriales y herramientas heterogéneas que no equivalen a una supervisión sistémica de la opacidad.

6.- Derechos fundamentales y remedios. El almacén europeo integra derechos fundamentales como criterios de diseño y despliegue, activando controles de legalidad (por ejemplo, límites a decisiones automatizadas, información significativa, supervisión humana). En EE. UU., la tutela se articula por leyes antidiscriminatorias y consumo, con remedios reparatorios y acuerdos que corrigen a posteriori (por ejemplo, cambios en sistemas, compensaciones), pero sin garantía estructural *ex ante* contra la opacidad.

7.- Extraterritorialidad y poder de mercado. La UE intenta proyectar sus estándares mediante el efecto Bruselas, imponiendo la conformidad a productos y servicios que se ofrecen en su mercado. EE. UU. proyecta su poder tecnológico y empresarial, condicionando de facto los parámetros de diseño global. La tensión entre normas europeas y arquitecturas tecnológicas mayoritariamente desarrolladas en EE. UU. mantiene bolsas de opacidad difíciles de auditar plenamente.

8.- Impacto práctico. En la UE, los desafíos residen en definir operativamente “inteligibilidad suficiente”, hacer auditables modelos complejos y dotar de medios a los supervisores. En EE. UU., el reto es superar la fragmentación y evitar que la transparencia dependa del litigio o de la voluntad empresarial. En ambos casos, la IA no “decide” por sí misma: la opacidad es resultado de elecciones humanas de diseño, datos y gobernanza; por eso el problema es jurídico e institucional antes que puramente técnico.

9.- Asimetría tecnológica y extraterritorialidad. La exigibilidad de garantías frente a la opacidad algorítmica se enfrenta a una asimetría estructural: gran parte de los sistemas usados en la UE se diseñan, entrenan o alojan fuera de su jurisdicción, mediante cadenas técnicas y contractuales —proveedores en la nube, subprocesadores, APIs o filiales— que diluyen la responsabilidad y dificultan el acceso a evidencias verificables. Aunque el RIA extiende sus obligaciones a los proveedores extraterritoriales que operan en el mercado europeo, su efectividad depende de que las autoridades y auditores puedan reconstruir el ciclo de vida del sistema —datos, trazabilidad del modelo, registros de inferencia y cambios de versión— y comprobar cómo se adoptan las decisiones en entornos reales.

Este desajuste entre estándar normativo y capacidad real de escrutinio se exagera por la invocación del secreto comercial, las restricciones de privacidad y seguridad, y la dispersión de evidencias técnicas entre múltiples actores. El resultado práctico es la persistencia de “zonas de opacidad” incluso bajo un marco ambicioso: los usuarios institucionales carecen de visibilidad suficiente para interpretar y justificar las salidas del sistema, y las autoridades ven limitada su posibilidad de reproducir, auditar y sancionar con base técnica sólida.

La extraterritorialidad, por tanto, no basta por sí sola: si no va acompañada de mecanismos efectivos de acceso (confidenciales y proporcionados) a documentación y evidencias técnicas, trazabilidad transfronteriza y coordinación regulatoria para evitar el arbitraje jurisdiccional, corre el riesgo de convertirse en una pretensión formal más que en una garantía material. En síntesis, el reto europeo no es solo declarar obligaciones globales, sino hacerlas verificables en un ecosistema técnico y empresarial que opera mayoritariamente fuera de sus fronteras [7].

Como conclusión, el modelo europeo institucionaliza la transparencia algorítmica como garantía estructural *ex ante*, mientras que el estadounidense la delega en gran medida a la autorregulación y a la aplicación caso a caso con carácter *ex post*. Ninguno de los dos garantiza por sí solo una protección plena frente a la opacidad: en la UE faltan definiciones y capacidades para una aplicación uniforme, y en EE. UU. faltan obligaciones federales que aseguren trazabilidad e inteligibilidad mínimas en todo el territorio. Esta brecha explica por qué, pese a marcos distintos, la opacidad persiste y sigue afectando a derechos como la igualdad, la privacidad y el debido proceso.

8. Propuestas normativas y principios comunes para combatir la opacidad algorítmica

La expansión de la decisión automatizada exige pasar de los valores a las garantías verificables. A continuación, se proponen unos principios operativos que pueden articularse tanto en la UE como en EE. UU., respetando sus culturas jurídicas, pero convergiendo en salvaguardas mínimas comunes.

Derecho a la inteligibilidad algorítmica (contenido mínimo exigible).

La inteligibilidad no consiste en explicar el algoritmo, sino en hacer comprensible la decisión a quien la experimenta. Supone que la persona afectada conozca que se emplea un sistema de inteligencia artificial, entienda su finalidad y sus límites, y pueda interpretar y cuestionar el resultado sin necesidad de conocimientos técnicos. El contenido mínimo debe incluir: notificación previa del uso de IA; descripción funcional clara (qué hace el sistema y qué no); explicación significativa del caso concreto (factores que más influyeron, cómo se combinaron y con qué grado de confianza); y un canal de asistencia humana para resolver dudas y ejercer derechos.

La proporcionalidad al riesgo determina la profundidad de la información: en ámbitos de alto impacto (empleo, vivienda, prestaciones, sanciones), la explicación debe ser más granular e incorporar acceso al expediente mínimo del caso (criterios aplicados, datos pertinentes, registros de la decisión) y vías de impugnación visibles. La inteligibilidad no exige revelar código fuente ni secretos industriales, sino ofrecer razones verificables y trazables, respaldadas por las evidencias técnicas internas y expresadas en lenguaje claro y formatos accesibles, evitando el exceso de documentación que oculte lo esencial. Sin inteligibilidad, la transparencia se vuelve

decorativa, con ella, la persona puede comprender, cuestionar y corregir decisiones automatizadas, y las instituciones pueden rendir cuentas de forma efectiva.

Trazabilidad y memoria técnica obligatoria.

La transparencia, para ser operativa, exige una memoria técnica capaz de reconstruir qué modelo tomó una decisión, con qué datos y bajo qué configuración. Por eso la trazabilidad debe abarcar todo el ciclo de vida: versionado de conjuntos de datos y modelos (incluida arquitectura y pesos), documentación de entrenamiento y validación (métricas y resultados por subgrupos) y registro de inferencia en producción (modelo y versión, marcas temporales, parámetros relevantes y, cuando proceda, elementos de explicación). Esta memoria debe ser íntegra y verificable (sellado temporal, huellas criptográficas, acceso por roles) y gestionarse con control de cambios y registro de procedencia que vincule datos, código y resultados.

La proporcionalidad al riesgo fija el nivel de detalle y los plazos de conservación: cuanto mayor sea el impacto en derechos, más granular y prolongada la retención. Para asegurar la comparabilidad, deben emplearse plantillas estandarizadas (*model cards* y *datasheets*) que resumen propósito, desempeño, límites y condiciones de uso [25-26]. En el sector público o cuasi público, la trazabilidad debe integrarse en registros institucionales, con una versión pública acompañada de un anexo confidencial para las autoridades. Sin memoria técnica, el sistema se vuelve inauditable e inimpugnable; con ella, es explicable, revisable y reparable.

Auditoría independiente proporcional al riesgo.

La auditoría es el mecanismo que convierte la transparencia en verificación. En usos de alto impacto, debe operar en dos tiempos: *ex ante* (antes del despliegue, para validar diseño, datos y controles) y *ex post* (de forma periódica, para comprobar desempeño real, deriva del modelo y cumplimiento). La realizan terceros acreditados e independientes, con reglas claras de incompatibilidades, ausencia de conflicto de interés y rotación de equipos.

El alcance de la auditoría abarca la documentación técnica, los conjuntos de datos y su trazabilidad, las configuraciones y el código relevantes, los registros de entrenamiento e inferencia, los resultados por subgrupos, así como los aspectos de seguridad y gobernanza. Para ello, el auditor necesita acceso supervisado a evidencias técnicas y muestras suficientes para garantizar la reproducibilidad, con medidas de confidencialidad (entornos seguros, acuerdos de confidencialidad y minimización de datos).

El equilibrio entre secreto comercial y control democrático se logra mediante acceso confidencial y proporcional para autoridades y auditores, y la publicación de un resumen ejecutivo que explique alcance, hallazgos, limitaciones y medidas correctoras sin exponer información sensible. La proporcionalidad guía la frecuencia y cobertura: cuanto mayor sea el riesgo, más intensiva la auditoría. En bajo riesgo, valen autoevaluaciones reforzadas y

muestreos aleatorios. Ante cualquier cambio de versión, indicio de degradación o incidente relevante, debe realizarse una nueva auditoría y activarse un plan de remediación con plazos definidos.

Contestación y reversibilidad efectivas.

La protección de derechos requiere que toda decisión automatizada sea impugnabile de forma real y rápida. Ello implica canales accesibles (formularios sencillos, sin coste, vías digital y presencial), plazos breves y perentorios y asistencia para personas no expertas. La revisión debe ser humana y cualificada, con potestad para apartarse del resultado algorítmico y con la obligación de motivar su decisión. Para que la impugnación sea efectiva, el afectado ha de acceder al expediente mínimo del caso: explicación significativa, datos y reglas aplicadas y registros pertinentes de la decisión.

En decisiones de alto impacto (prestaciones, vivienda, sanciones, medidas que limitan derechos), el efecto suspensivo debe ser la regla mientras se revisa, salvo urgencia motivada con medidas provisionales que minimicen daños. Los remedios no se agotan en la mera resolución del caso: incluyen la rectificación, un nuevo cálculo (*re-score*) con la versión corregida, la retroacción de actuaciones, la compensación cuando proceda y las medidas estructurales necesarias (ajustes del sistema, reentrenamiento, desactivación de versiones defectuosas, notificación a terceros afectados) bajo el principio de no repetición.

La proporcionalidad al riesgo guía el diseño: en bajo impacto bastan respuestas ágiles y explicaciones reforzadas. En alto impacto, procedimientos formales con plazos máximos garantizados y métricas públicas (tiempos de respuesta, tasa de revocaciones). Sin contestación ni reversibilidad, la transparencia pierde eficacia, con ellas, la automatización se vuelve corregible y responsable.

Monitorización posdecpliegue y gestión de la deriva

La calidad de un sistema no se congela al desplegarlo: cambia con el entorno, los datos y los usos. Por eso se requiere un seguimiento continuo que detecte la deriva de los datos (cambios en la distribución de las entradas) y la deriva del modelo (pérdida de rendimiento o variaciones en las relaciones relevantes). La supervisión continua debe cubrir métricas de desempeño y de equidad por subgrupos, con alertas automáticas ante degradaciones, umbrales predefinidos y responsables claros de respuesta.

Cuando se superen los umbrales de riesgo, deben activarse medidas graduales de mitigación: recalibración, reentrenamiento controlado, retorno a una versión estable, intervención humana reforzada o detención segura del sistema hasta restaurar condiciones aceptables. Todo cambio debe someterse a una gestión documentada, que incluya la justificación, las pruebas en entorno controlado (por ejemplo, ejecuciones paralelas o pruebas piloto), la

evaluación de impacto y el registro de procedencia técnica que vincule datos, configuración y resultados.

La proporcionalidad al riesgo determina la frecuencia y profundidad del seguimiento: en sistemas de alto impacto, deben existir paneles de control con vigilancia diaria, revisión por pares y reportes periódicos; en los de menor impacto, bastan muestreos y revisiones programadas. Asimismo, debe establecerse un plan de retirada o sustitución del modelo, que contemple el calendario de desactivación, la continuidad operativa y la notificación a las personas afectadas y a las autoridades competentes cuando proceda. Sin supervisión y gestión de la deriva, un sistema que fue válido puede perder su fiabilidad con el paso del tiempo, sin que nadie lo advierta a tiempo.

Métricas de equidad y desempeño por subgrupos.

Para saber si un sistema trata de forma consistente a personas situadas en contextos distintos, no basta con una métrica global: se necesitan mediciones desagregadas por subgrupos relevantes (por ejemplo, sexo, edad, origen racial/étnico, discapacidad u otros legalmente pertinentes), definidos caso a caso y con garantías de protección de datos. En alto riesgo, el reporte debe ser periódico (por ejemplo, trimestral) y compararse con una línea base y con los resultados del periodo anterior.

El informe debería incluir, al menos, las tasas de falsos positivos y negativos, el error de calibración (grado en que las puntuaciones reflejan la realidad), las métricas por subgrupos y la estabilidad de las explicaciones (si la razón ofrecida varía de forma arbitraria ante perturbaciones mínimas). Cuando existan disparidades significativas, deben justificarse los compromisos entre objetivos: cuál es el criterio de equidad adoptado (por ejemplo, equilibrio de errores o calibración entre grupos), por qué resulta adecuado al contexto y qué medidas correctoras se aplicarán. Las metas de equidad han de ser razonadas y medibles (con rangos o umbrales definidos), evitando imponer paridades matemáticas indiscriminadas que puedan generar un trato desigual no justificado. En consecuencia, conviene definir límites y márgenes de disparidad que incentiven la mejora continua en lugar de imponer uniformidad artificial.

Los resultados de este seguimiento deben comunicarse de manera transparente: mediante un resumen ejecutivo público —que exponga la metodología, las métricas clave, los principales hallazgos y las acciones emprendidas— y, cuando proceda, un anexo confidencial con los detalles técnicos y los recuentos mínimos necesarios para proteger la privacidad.

Registro público y transparencia institucional.

En el sector público y en entidades cuasi públicas o concesionarias, la transparencia debe materializarse en un registro de sistemas algorítmicos accesible y actualizado. El registro debería identificar, como mínimo: el propósito y la base jurídica; las categorías y fuentes de

datos (incluidas las categorías especiales, cuando proceda); el proveedor y la cadena de subprocesadores; la versión del sistema y la fecha de despliegue; las métricas clave y las limitaciones conocidas; los resultados de las auditorías (resumen ejecutivo) y las evaluaciones de impacto en derechos en su versión pública; así como el responsable del sistema y el punto de contacto para reclamaciones.

El registro debe ser un instrumento vivo: toda actualización relevante (cambio de versión, reentrenamiento, incorporación de nuevos flujos de datos) debe anotarse y fecharse, conservando un historial completo que garantice su trazabilidad. Su consulta ha de ser clara y legible, en formatos abiertos y, cuando sea posible, accesible mediante interfaces programáticas (APIs) que permitan su reutilización por investigadores y organizaciones de la sociedad civil. La información sensible podrá reservarse en un anexo confidencial destinado a las autoridades y auditores, asegurando el equilibrio entre el control democrático y la confidencialidad empresarial o de seguridad.

En las interfaces que interactúan directamente con las personas —formularios, notificaciones o portales— debe mostrarse un etiquetado visible de “decisión asistida por IA”, con un enlace al registro que explique su alcance, los datos utilizados y las vías de impugnación disponibles. La existencia de un registro público no solo previene la opacidad, sino que convierte la transparencia en una forma efectiva de rendición de cuentas verificable.

Incidentes e investigación obligatoria.

Para que la gestión del riesgo sea creíble, hay que definir qué es un “incidente grave”: daño a derechos (por ejemplo, denegación injusta de una prestación), brechas de seguridad o privacidad, discriminación material por subgrupos, fallos sistémicos o uso fuera de propósito. Ante un incidente, se debe exigir notificación temprana a la autoridad competente y, cuando proceda, comunicación a las personas afectadas con lenguaje claro (qué ocurrió, a quién afecta, qué derechos asisten y cómo ejercerlos).

Debe abrirse una investigación independiente y documentada, con acceso supervisado a evidencias técnicas (registros, versiones, configuraciones) y preservación de la cadena de custodia. Las primeras 24–72 horas se deberían orientar a la contención (desactivar o degradar el sistema, restaurar la última versión estable, intervención humana reforzada) y a recolección de pruebas. El cierre debería exigir un plan de corrección (parches, reentrenamiento, cambios de datos/procesos, controles adicionales) y un informe posincidente con análisis de causa raíz, lecciones aprendidas y medidas de no repetición. Debe publicarse un resumen ejecutivo sin datos sensibles y actualizar el registro de sistemas y el mapa de riesgos. La proporcionalidad guía la profundidad de la investigación y la amplitud de la notificación. La trazabilidad y la retención de registros garantizan una auditoría posterior. Sin este ciclo de notificar–investigar–corregir–aprender, la transparencia no protege efectivamente.

Contratación pública y cumplimiento por contrato.

La contratación es la palanca más eficaz para convertir principios en obligaciones exigibles: *sin cláusula, no hay despliegue*. En contratos del sector público (y de entidades reguladas) deben incorporarse, como mínimo: derechos de auditoría con acceso confidencial y proporcional para autoridades y auditores, entrega periódica de evidencias técnicas (fichas del modelo y del conjunto de datos, documentación de entrenamiento y validación, elementos de explicabilidad por caso y métricas por subgrupos), límites de finalidad y reutilización de datos (minimización, prohibición de usos secundarios no autorizados), retención proporcional y borrado verificable.

Para garantizar la continuidad y el control, deben incluirse cláusulas de portabilidad (formato abierto, documentación para la migración) y un depósito en garantía regulatoria: un almacenamiento cifrado del modelo, sus pesos, configuraciones, documentación y claves de reconstrucción, custodiado por un tercero o por la autoridad competente, y accesible en caso de investigación, cese del servicio o incumplimiento grave. Las obligaciones deben transmitirse a los subprocesadores mediante cláusulas de extensión contractual, y preverse la gestión de incidentes con plazos máximos garantizados, plan de remediación y reporte.

Como elemento de cierre indispensable, el contrato debe prever un régimen sancionador y de resolución (penalizaciones, suspensión, rescisión por incumplimiento), junto con garantías financieras y compromisos sobre propiedad intelectual que no bloqueen el acceso a las evidencias técnicas necesarias para la supervisión. Así, el cumplimiento por vía contractual evita que el secreto comercial actúe como un escudo absoluto y hace operativa la rendición de cuentas antes y después del despliegue.

Gobernanza organizativa y rendición de cuentas.

La transparencia solo se convierte en garantía cuando existe responsabilidad clara dentro de la organización. Para ello, conviene designar una figura responsable de IA con independencia funcional, presupuesto y acceso directo al consejo, capaz de frenar despliegues si hay riesgos no mitigados. Esta figura debe coordinar un comité interno multidisciplinar (datos, jurídico, negocio, ética, ciberseguridad) que aprueba políticas de IA, define roles y responsabilidades y mantiene un registro de riesgos con controles y propietarios.

La formación obligatoria para desarrolladores, responsables de decisión y personal operativo constituye un pilar esencial del control interno: no debe limitarse a los aspectos técnicos, sino incluir también formación en derechos fundamentales, no discriminación y responsabilidad ética. El sistema de gobernanza debe complementarse con una segunda línea de cumplimiento, encargada de la supervisión continua, y una tercera línea de auditoría interna que evalúe la eficacia de los controles, garantizando canales de escalado e independencia frente a posibles conflictos de interés.

Anualmente, la organización emitirá una atestación de conformidad basada en un marco de gestión de riesgos (por ejemplo, NIST AI RMF 1.0) con evidencias trazables: metas, métricas por subgrupos, resultados de auditorías, incidentes y acciones correctoras [26]. El consejo debe recibir informes periódicos y aprueba planes de remediación. Sin esta arquitectura de gobernanza y rendición de cuentas, las obligaciones técnicas quedan en declaraciones, con ella, pasan a ser compromisos verificables.

Protección de denuncias y equilibrio con secreto comercial.

Para detectar a tiempo riesgos algorítmicos, la organización necesita canales seguros de denuncia accesibles a personal interno y proveedores (incluida la opción anónima), con confidencialidad garantizada y prohibición de represalias. Estos canales deben permitir comunicar incidentes, sesgos, incumplimientos o manipulaciones y activar una revisión independiente con plazos máximos y trazabilidad de la respuesta.

La divulgación responsable a autoridades requiere de puertos seguros: las personas alertadoras que actúen de buena fe y con base razonable quedan protegidas, aunque existan cláusulas de confidencialidad. El secreto comercial no puede operar como escudo absoluto: se deben prever rutas de acceso confidencial y proporcional a la evidencia técnica (documentación, registros, configuraciones, modelos) para supervisores y auditores públicos.

Para mantener el equilibrio, el marco debe contemplar sanciones proporcionadas ante filtraciones maliciosas o uso indebido de información, sin penalizar la denuncia legítima ni disuadirla. En paralelo, hay que ofrecer orientación jurídica y apoyo a quien informe, y garantizar retroalimentación sobre las medidas adoptadas. Esta arquitectura convierte la denuncia en un mecanismo de prevención y corrección, preservando la integridad del secreto comercial cuando sea pertinente, pero subordinándolo al interés público y a la tutela efectiva de derechos.

Convergencia técnica y reconocimiento mutuo.

Para que la gobernanza sea operativa a escala global, conviene alinear prácticas técnicas y reconocer certificaciones cuando exista equivalencia sustantiva. El punto de encuentro puede ser un lenguaje común de riesgos y controles (por ejemplo, taxonomías compatibles con el mencionado NIST AI RMF 1.0) y plantillas estándar (fichas del modelo y fichas del conjunto de datos) que documenten propósito, límites, métricas y condiciones de uso de manera comparable [24–26]. Con ello se reduce el coste de cumplimiento, se facilita la portabilidad de evidencias (mismas categorías y formatos) y se evita el arbitraje regulatorio. El reconocimiento mutuo no debe ser un “mínimo común” sino una equivalencia demostrada: misma cobertura de riesgo, profundidad de auditoría, competencia del organismo evaluador y seguimiento periódico (vigilancia poscertificación). Ayuda establecer mapeos públicos entre marcos (qué

control del país A satisface el del país B), catálogos de controles interoperables y formatos legibles por máquina para intercambiar registros de evaluación.

Ajustes jurisdiccionales.

La eficacia de las garantías depende de cómo las aplica cada jurisdicción. En la UE, conviene reforzar un “mínimo común operativo”: (i) directrices técnicas comunes que traduzcan obligaciones listas de verificación auditables; (ii) plantillas de auditoría y de memoria técnica (qué evidencias deben conservarse y por cuánto tiempo) (iii) guías sectoriales de “inteligibilidad suficiente” (empleo, vivienda, salud, justicia), con ejemplos de explicaciones válidas e inválidas y (iv) capacidades de supervisión: equipos técnicos mixtos, bolsas de expertos, laboratorios de prueba y cooperación entre autoridades. Esto reduce la disparidad entre Estados miembros y facilita el reconocimiento intra-UE de evaluaciones. Por su parte, en EE. UU., el reto es establecer un marco federal mínimo (transparencia, trazabilidad y contestabilidad) compatible con las competencias estatales, para evitar protecciones desiguales y la dependencia del litigio como vía principal de corrección. Ese suelo federal puede articular estándares de información, retención de evidencias técnicas, procedimientos de impugnación y reporte por subgrupos, dejando a los estados elevar el listón. Requiere además una coordinación interagencias (FTC, EEOC, CFPB, HUD, HHS) y modelos contractuales para compras públicas.

En ambos entornos, estos ajustes operativos acortan la distancia entre norma y práctica, y convierten principios en controles repetibles y auditables.

Lenguaje claro y accesibilidad.

La inteligibilidad solo se materializa si la información se presenta de forma comprensible y accesible. Las explicaciones deberán: (i) usar lenguaje claro (evitar jerga, definir términos), con estructura en capas (resumen breve + detalle ampliado) (ii) cumplir con la accesibilidad digital (por ejemplo, criterios equivalentes a WCAG 2.1 AA: contraste, navegación por teclado, textos alternativos, subtítulo) (iii) ofrecer formatos alternativos cuando sea necesario (lectura fácil, audio, intérprete, canales no digitales) y traducciones razonables en contextos multilingües (iv) evitar patrones oscuros y diseños que dificulten el ejercicio de derechos y (v) proporcionar puntos de contacto y tiempos máximos de respuesta para dudas y reclamaciones. Para facilitar la supervisión y la reutilización, deberá publicarse un resumen ejecutivo y, cuando proceda, datos y documentos en formatos legibles por máquina (por ejemplo, JSON o CSV para métricas), preservando siempre la privacidad. La calidad comunicativa se evaluará mediante métricas de legibilidad y pruebas con usuarios reales. De este modo, el derecho a la explicación se convierte en un derecho a la comprensión efectiva para todas las personas.

Con estas propuestas, la transparencia algorítmica deja de ser un freno y se convierte en condición de legitimidad democrática: permite trazar, entender, auditar y corregir sistemas que

afectan derechos, y distribuye responsabilidades entre proveedores, usuarios institucionales, auditores y autoridades. La clave no es solo publicar normas, sino hacerlas operables con pruebas, métricas, registros y mecanismos de reparación.

Conclusión

La opacidad algorítmica es hoy un problema estructural de las sociedades digitalizadas: no se agota en la complejidad técnica de los modelos, sino que emerge de elecciones humanas de diseño, de asimetrías informativas y de barreras jurídico-organizativas (secreto comercial, contratos opacos, falta de documentación y de control independiente) que impiden trazar, entender y discutir decisiones automatizadas con efectos reales sobre la igualdad, la no discriminación, la privacidad y el debido proceso. La inteligencia artificial no decide por sí misma, potencia y automatiza lógicas institucionales ya existentes. Precisamente por ello, el desafío es jurídico e institucional: cómo imponer condiciones de inteligibilidad y rendición de cuentas a tecnologías que median decisiones que antes eran plenamente humanas.

La comparación entre UE y EE. UU. evidencia dos racionalidades regulatorias. La UE intenta internalizar la transparencia como obligación *ex ante* ligada al riesgo, integrando documentación, supervisión humana e información significativa en un armazón común. EE. UU., por el contrario, opera con un mosaico sectorial y territorial que relega la contestación de la opacidad a litigios y a la autorregulación. Esta divergencia se agrava por la asimetría tecnológica: buena parte de los sistemas usados en Europa se diseñan fuera de su jurisdicción, tensionando la aplicación efectiva de sus estándares y el acceso a las evidencias técnicas necesarias para auditar. En ambos modelos persiste, además, el conflicto entre secreto comercial y escrutinio público.

Ninguno de los dos entornos ofrece todavía una protección efectiva y homogénea frente a la opacidad. En la UE, la distancia entre norma y práctica se explica por indeterminaciones operativas (qué es “inteligibilidad suficiente”), capacidades institucionales desiguales y dificultades para auditar modelos complejos en producción. En EE. UU., la ausencia de un suelo federal mínimo perpetúa protecciones desiguales y hace depender los remedios de acuerdos y demandas puntuales. En ambos casos, la transparencia no puede confundirse con publicar código o con descargar documentos: si no produce comprensión efectiva para el afectado, no cumple su función de garantía.

De ahí que la transparencia, junto con la trazabilidad, la auditabilidad y la contestabilidad, deba consolidarse como principio jurídico vinculante y operativo, con obligaciones verificables: memoria técnica y *logs* de decisiones, auditoría independiente proporcional al riesgo, métricas por subgrupos, seguimiento posdespliegue y vías de impugnación con efectos reales. Este desplazamiento desde la retórica de principios hacia mecanismos exigibles requiere

instituciones con capacidad técnica y autonomía, y una convergencia mínima de estándares que desactive el arbitraje regulatorio.

En suma, la legitimidad de la gobernanza algorítmica depende de convertir la transparencia en condición de posibilidad del Estado de derecho en entornos automatizados. Sin explicaciones comprensibles, huellas trazables y controles externos no hay derechos que puedan ejercerse ni tutelarse. Con ellos, y solo con ellos, la innovación tecnológica deja de ser una coartada para la opacidad y se reencuentra con su horizonte democrático.

Nota metodológica final

Este trabajo se ha elaborado mediante una metodología jurídico-crítica basada en tres pilares: (i) análisis normativo del marco europeo y estadounidense (Carta DFUE, RGPD, Directiva 2016/680 y Reglamento (UE) 2024/1689, junto con normas estatales y guías federales en EE. UU.); (ii) revisión doctrinal e informes institucionales sobre transparencia, trazabilidad e inteligibilidad, distinguiendo entre explicabilidad *post hoc*, interpretabilidad *ex ante* y transparencia procedimental; y (iii) estudio de casos paradigmáticos de opacidad algorítmica y su impacto en derechos fundamentales (SyRI, iBorderCtrl, iTutorGroup/AEDT, SafeRent, algoritmos sanitarios, Clearview AI y el acuerdo DOJ–Meta). El enfoque comparado UE–EE. UU. se articula en dos fases: descripción funcional de obligaciones y controles (*ex ante/ex post*), seguida de una evaluación crítica de su eficacia. La selección de casos combina diversidad sectorial (justicia, fronteras, empleo, vivienda, salud) y heterogeneidad regulatoria (judicialización, control administrativo y acuerdos de cumplimiento). Para mitigar sesgos de confirmación, se contrastaron fuentes académicas con documentos oficiales y resoluciones judiciales, priorizando materiales verificables. La propuesta normativa (apartado 8) traduce los principios en salvaguardas operativas —trazabilidad, auditoría, contestación, métricas y seguimiento— diferenciándolas de la comparación crítica del apartado 7.

Limitaciones del estudio. (i) Dinamismo regulatorio: el marco normativo y la jurisprudencia evolucionan con rapidez, algunas obligaciones o criterios interpretativos pueden haberse actualizado tras la fecha de cierre de este trabajo. (ii) Acceso restringido a evidencias técnicas y contractuales: por razones de secreto comercial, protección de datos y seguridad, no siempre es posible acceder al código, pesos de los modelos, conjuntos de entrenamiento, documentación interna, contratos y registros de producción. Esta falta de acceso puede limitar la verificación externa plena y la cuantificación inmediata de impactos en “tiempo real”.

Descargo de responsabilidad

Este trabajo se ha producido en el contexto de la Cátedra en Inteligencia Artificial en Banca, financiada por Rural Servicios Informáticos (RSI). El contenido refleja exclusivamente las

opiniones de los autores. RSI y la Universidad de Alcalá no se hacen responsables de las ideas, análisis o conclusiones aquí expresadas, ni de su eventual utilización por terceros.

Referencias

1. BAROCAS, S. y SELBST, A. D. *Big Data's Disparate Impact*. California Law Review [en línea]. 2016, vol. 104, n.º 3, pp. 671-732. DOI: 10.2139/ssrn.2477899 [Acceso: 1 octubre 2025].
2. EUROPEAN UNION. *Directive (EU) 2016/680...* Official Journal of the European Union [en línea]. 2016, L 119, pp. 89–131. Disponible en: <https://eur-lex.europa.eu/eli/dir/2016/680/oj> [Acceso: 1 octubre 2025].
3. EUROPEAN UNION. *Regulation (EU) 2016/679 (General Data Protection Regulation)*. Official Journal of the European Union [en línea]. 2016, L 119, pp. 1–88. Disponible en: <https://eur-lex.europa.eu/eli/reg/2016/679/oj> [Acceso: 1 octubre 2025].
4. EUROPEAN UNION. *Charter of Fundamental Rights of the European Union*. Official Journal of the European Union [en línea]. 2012, C 326, pp. 391–407. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT> [Acceso: 1 octubre 2025].
5. UNITED NATIONS. *International Covenant on Civil and Political Rights*. New York: United Nations, 1966 [en línea]. Disponible en: <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights> [Acceso: 1 octubre 2025].
6. MANTELERO, A. *AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment*. Computer Law & Security Review [en línea]. 2018, 34(4), 754-772. DOI: 10.1016/j.clsr.2018.05.017 [Acceso: 1 octubre 2025].
7. EUROPEAN UNION. *Regulation (EU) 2024/1689... on Artificial Intelligence*. Official Journal of the European Union [en línea]. 2024, L 202, pp. 1-127. Disponible en: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689> [Acceso: 1 octubre 2025].
8. EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS (FRA). *Getting the Future Right – Artificial Intelligence and Fundamental Rights*. Luxembourg: Publications Office of the EU [en línea]. 2020. Disponible en: <https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights> [Acceso: 1 octubre 2025].
9. RECHTBANK DEN HAAG. *NJCM et al. v The Dutch State (SyRI)*, ECLI:NL:RBDHA:2020:1878. Uitspraken – Rechtspraak.nl [en línea]. 5 febrero 2020. Disponible en: <https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:1878>. PDF: https://www.escri-net.org/wp-content/uploads/2020/09/ecli_nl_rbdha_2020_1878.pdf [Acceso: 1 octubre 2025].
10. ZUIDERVEEN BORGESIOUS, F. J. *A Note on the SyRI Judgment*. Tilburg University [en línea]. 2020. Disponible en: https://research.tilburguniversity.edu/files/43647493/syri_case_note.pdf [Acceso: 1 octubre 2025].

11.COURT OF JUSTICE OF THE EUROPEAN UNION. *Case T-158/19, Breyer v European Commission*. CURIA [en línea]. 15 diciembre 2021. Disponible en: <https://curia.europa.eu/juris/document/document.jsf?docid=277067&doclang=EN> [Acceso: 1 octubre 2025].

12.EUROPEAN DIGITAL RIGHTS (EDRi). *European Court supports transparency in risky EU border tech experiments*. EDRi [en línea]. 16 diciembre 2021. Disponible en: <https://edri.org/our-work/european-court-supports-transparency-in-risky-eu-border-tech-experiments/> [Acceso: 1 octubre 2025].

13.CRAWFORD, K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press, 2021. ISBN 9780300209570.

14.STATE OF ILLINOIS. *Biometric Information Privacy Act (BIPA), 740 ILCS 14*. Illinois Compiled Statutes [en línea]. 2008—. Disponible en: <https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57> [Acceso: 1 octubre 2025].

15.NEW YORK CITY DEPARTMENT OF CONSUMER AND WORKER PROTECTION (DCWP). *Automated Employment Decision Tools (AEDT)*. Página oficial [en línea]. Disponible en: <https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page>. Texto legal codificado: <https://codelibrary.amlegal.com/codes/newyorkcity/latest/NYCadmin/0-0-0-135598>. FAQ (PDF): <https://www.nyc.gov/assets/dca/downloads/pdf/about/DCWP-AEDT-FAQ.pdf> [Acceso: 1 octubre 2025].

16.FEDERAL TRADE COMMISSION. *The Federal Trade Commission 2023 Privacy and Data Security Update* [en línea]. 21 marzo 2024. Disponible en: https://www.ftc.gov/system/files/ftc_gov/pdf/2024.03.21-PrivacyandDataSecurityUpdate-508.pdf. (Apoyo: *Aiming for truth, fairness, and equity in your company's use of AI*, 19 abril 2021, PDF) Disponible en: <https://privacysecurityacademy.com/wp-content/uploads/2021/04/Aiming-for-truth-fairness-and-equity-in-your-companys-use-of-AI.pdf> [Acceso: 1 octubre 2025].

17.ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L. *Machine Bias*. ProPublica [en línea]. 23 mayo 2016. Disponible en: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [Acceso: 1 octubre 2025].

18.U.S. EQUAL EMPLOYMENT OPPORTUNITY COMMISSION. *iTutorGroup to Pay \$365,000 to Settle EEOC Discriminatory Hiring Suit* [en línea]. 9 agosto 2023. Disponible en: <https://www.eeoc.gov/newsroom/itutorgroup-pay-365000-settle-eeoc-discriminatory-hiring-suit> [Acceso: 1 octubre 2025].

19.GREATER BOSTON LEGAL SERVICES; COHEN MILSTEIN y otros. *Rental Applicants Using Housing Vouchers Settle Ground-Breaking Discrimination Class Action Against SafeRent Solutions* (nota de prensa) [en línea]. 26 abril 2024. Disponible en:

<https://www.gbls.org/sites/default/files/2024-04/SafeRent-press-release-settlement-reached-4-26-2024.pdf>
[Acceso: 1 octubre 2025].

20.ASSOCIATED PRESS. *Class action lawsuit on AI-related discrimination reaches final settlement* [en línea]. 20 noviembre 2024. Disponible en: <https://apnews.com/article/1bc785c24a1b88bd425a8fa367ab2b23> [Acceso: 1 octubre 2025].

21.OBERMEYER, Z.; POWERS, B.; VOGELI, C.; MULLAINATHAN, S. *Dissecting racial bias in an algorithm used to manage the health of populations*. Science [en línea]. 2019, 366(6464), 447–453. DOI: [10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342) [Acceso: 1 octubre 2025].

22.AMERICAN CIVIL LIBERTIES UNION. *ACLU v. Clearview AI* [en línea]. 2022– . Disponible en: <https://www.aclu.org/cases/aclu-v-clearview-ai> [Acceso: 1 octubre 2025].

23.U.S. DEPARTMENT OF JUSTICE. *Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Inc...* [en línea]. 21 junio 2022. Disponible en: <https://www.justice.gov/archives/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known> [Acceso: 1 octubre 2025].

24.NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1 [en línea]. 2023. Disponible en: <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf> (overview: <https://www.nist.gov/itl/ai-risk-management-framework>) [Acceso: 1 octubre 2025].

25.MITCHELL, M.; WU, S.; ZALDIVAR, A.; BARNES, P.; VASSERMAN, L.; HUTCHINSON, B.; SPITZER, E.; RAJI, I. D.; GEBRU, T. *Model Cards for Model Reporting*. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019) [en línea]. 2019. DOI: [10.1145/3287560.3287596](https://doi.org/10.1145/3287560.3287596). Preprint: <https://arxiv.org/abs/1810.03993> [Acceso: 1 octubre 2025].

26.GEBRU, T., et al. *Datasheets for Datasets*. Communications of the ACM [en línea]. 2021, 64(12), 86–92. DOI: [10.1145/3458723](https://doi.org/10.1145/3458723). Preprint: <https://arxiv.org/abs/1803.09010> [Acceso: 1 octubre 2025].

Nacho Hernandez Moreno. La Inteligencia Artificial ¿Una distopía para los Derechos Humanos?

Nacho Hernandez Moreno

Doctor en Derechos Humanos, democracia y justicia internacional
Profesor ayudante doctor, Departamento de Filosofía del Derecho y Política, Facultad de
Derecho, y miembro del Institut de Drets Humans de la Universitat de València

[ACCEDER AL VIDEO](#)

Resumen

La expansión de la inteligencia artificial («IA») está transformando de forma decisiva los vínculos entre conocimiento, poder y libertad, con implicaciones directas para la deliberación democrática y los derechos humanos. Este trabajo analiza cómo la automatización de la información y la mediación algorítmica reconfiguran la deliberación pública y los derechos humanos. Tomando como punto de partida la afirmación de López de Mántaras —«sin verdad verificable no hay deliberación posible y, sin deliberación, la democracia pierde sentido»—, se examina la IA como infraestructura social y política que condiciona el acceso al conocimiento, la visibilidad del discurso y la distribución del poder informacional. El estudio explora tres ilusiones centrales de la tecnología —autonomía, transparencia y neutralidad— y muestra cómo la opacidad algorítmica y la personalización extrema generan un ecosistema comunicativo propenso a la manipulación y a la desigualdad. El tránsito del ideal ciberutópico al capitalismo de vigilancia descrito por Zuboff ha consolidado un modelo económico basado en la extracción de datos y la predicción conductual, que amenaza la libertad de expresión, la privacidad y la igualdad. A través del diálogo con autores como Habermas, Dahl, Floridi y Sunstein, se argumenta que la IA no sustituye la deliberación democrática, pero sí puede vaciarla de contenido si los algoritmos reemplazan la racionalidad por la eficiencia y la predicción. El trabajo concluye proponiendo una doble respuesta ética y política: una ética del diseño responsable, fundada en los principios de justicia y explicabilidad, y una política de control democrático del poder informacional mediante transparencia, alfabetización digital y redistribución de la soberanía tecnológica. Solo así la IA podrá convertirse en un instrumento al servicio de la libertad y no en una distopía para los derechos humanos.

Palabras clave: inteligencia artificial, democracia, derechos humanos, capitalismo de vigilancia, deliberación pública.

Abstract

The expansion of artificial intelligence (AI) is reshaping the dynamics linking knowledge, power, and freedom, which direct implications for democratic deliberation and human rights protection. This paper examines how information automation and algorithmic mediation transform public deliberation and human rights. Building on López de Mántaras's assertion—"without verifiable truth there can be no deliberation, and without deliberation democracy loses its meaning"—AI is analyzed as a social and political infrastructure that structures access to knowledge, discourse visibility, and the distribution of informational power. The study explores three central illusions of technology—autonomy, transparency, and neutrality—and shows how algorithmic opacity and extreme personalization foster manipulation and inequality. The shift from the early cyber-utopian ideal to Zuboff's surveillance capitalism has created an economy based on data extraction and behavioral prediction, threatening freedom of expression, privacy, and equality. Drawing on authors such as Habermas, Dahl, Floridi, and Sunstein, the paper argues that AI does not abolish democratic deliberation but can render it hollow when algorithms replace reason with efficiency and prediction. It concludes by proposing an ethical and political response: an ethics of responsible design grounded in justice and explainability, and a democratic governance of informational power based on transparency, digital literacy, and redistribution of technological sovereignty. Only under these conditions can AI serve freedom instead of becoming a dystopia for human rights.

Key words: artificial intelligence, democracy, human rights, surveillance capitalism, public deliberation.

1. Introducción

La inteligencia artificial («IA») se ha convertido en una tecnología estructural del siglo XXI. Se halla presente en una multiplicidad de ámbitos de nuestra vida cotidiana. Por ejemplo, se emplea para buscar información, traducir textos, recibir recomendaciones de libros o películas, gestionar nuestras agendas, desbloquear nuestros dispositivos por reconocimiento facial, encontrar la ruta más rápida o sostenible para llegar a nuestro destino o para filtrar correos electrónicos no deseados. No obstante, pese a su potencial transformador, la IA plantea una serie de riesgos éticos, sociales y políticos que amenazan los valores fundacionales de la democracia y los derechos humanos. En el marco de esta investigación, analizaremos los problemas que implican la pérdida de privacidad y la vigilancia constante, la manipulación de la información y la erosión de la deliberación pública, la discriminación algorítmica y reproducción de sesgos, así como la concentración del poder tecnológico en manos de corporaciones que han dado vida al capitalismo de vigilancia descrito por Zuboff (2009).

Una sola frase sintetiza la relevancia de la investigación en torno a la inteligencia artificial y su influencia en la democracia y los derechos humanos: «sin verdad verificable, no hay deliberación posible y, sin deliberación, la democracia pierde sentido» (López de Mántaras, 2025). Estas palabras condensan de manera precisa la encrucijada en que se encuentran hoy las democracias contemporáneas. La IA, en particular los algoritmos de recomendación y clasificación que estructuran la comunicación digital como la personalización y la burbuja de filtros (*filter bubble*), ha alterado de manera radical la forma en que la ciudadanía accede a la información, participa en la esfera pública y ejerce sus derechos fundamentales.

Este trabajo tiene como objetivo principal examinar cómo la IA, a través de los algoritmos en el ciberespacio, puede socavar las condiciones mismas de posibilidad de la democracia y obstaculizar el ejercicio efectivo de los derechos humanos de la ciudadanía. Para ello, el presente documento comienza con una breve explicación en torno al concepto y la naturaleza de la IA para analizar posteriormente cómo se inserta en el capitalismo digital y el impacto que esta relación tiene en la democracia y en los derechos humanos.

2. Una aproximación a la inteligencia artificial

La IA puede definirse como «una colección de componentes computacionales que permiten construir sistemas que emulan funciones realizadas por el cerebro humano» (López de Mántaras & Brunet, 2024, p. 13). Estos sistemas son capaces de resolver problemas y ejecutar tareas de razonamiento, aprendizaje o percepción, aunque lo hagan sin comprensión genuina de lo que realizan. Su eficacia se basa en algoritmos y en el análisis masivo de datos, mediante los cuales se crean modelos que detectan patrones, infieren correlaciones y formulan predicciones. No constituye una forma de inteligencia propiamente dicha, sino procesamiento automatizado de información dirigido a un fin concreto.

A diferencia de la inteligencia humana, la IA no opera desde la experiencia ni desde el sentido común, sino desde la inferencia estadística. Por ello, puede reconocer un rostro con gran precisión, pero no entender una emoción; puede traducir textos, pero sin comprender los matices culturales; o puede recomendar contenido, pero sin saber qué consecuencias éticas o sociales se derivan de ello. López de Mántaras (2022, p. 5) señala que los sistemas actuales de IA carecen de «sentido común», el componente esencial que permite al ser humano interpretar lo que percibe y adaptar sus respuestas a lo imprevisto. Por tanto, los sistemas de IA no piensan ni entienden: simplemente calculan. De hecho, la IA es eficaz precisamente en la ejecución de tareas para las cuales nuestra inteligencia no es necesaria: «es precisamente cuando dejamos de intentar producir inteligencia humana cuando podemos empezar a sustituirla con notable éxito en un número cada vez mayor de tareas» (Floridi, 2024, p. 83). Como destaca este autor, más que entidades artificiales *inteligentes*, son *agentes* artificiales capaces de actuar en el entorno de manera autónoma dentro de unos límites predefinidos. Son entidades informacionales que operan, interactuando, aprendiendo y adaptando su comportamiento, pero sin consciencia.

La confusión entre inteligencia y agencia, entre comprender y ejecutar, explica buena parte de las expectativas desmesuradas y de los temores distópicos que acompañan a la IA. En realidad, los sistemas inteligentes actúan dentro de límites estrictamente definidos, siguiendo objetivos y parámetros establecidos por seres humanos. Sin embargo, su creciente sofisticación, su integración en procesos sociales y su capacidad para tomar decisiones con consecuencias tangibles han generado una serie de ilusiones —de autonomía, de transparencia y de neutralidad— que distorsionan nuestra comprensión de la IA y sus implicaciones para la democracia y los derechos humanos.

2.1. La ilusión de autonomía

La idea de que las máquinas son o serán verdaderamente autónomas, presentadas como entes que aprenden y actúan sin intervención humana e incluso como potenciales sujetos decisorios es una noción engañosa. Los sistemas de IA pueden funcionar de manera automática, pero no autónoma. En puridad, su aparente autonomía es una de carácter operativo o funcional, que permite a un sistema ejecutar tareas sin intervención humana inmediata, pero no de índole moral, exclusiva de agentes conscientes. Esta última implica voluntad, comprensión y capacidad para deliberar sobre los objetivos, pero las máquinas solo pueden ejecutar los medios dentro de un marco de metas y reglas codificadas o aprendidas. Los algoritmos carecen de fines propios: optimizan funciones que les han sido asignadas. En consecuencia, su aparente independencia no es sino una delegación humana de decisiones.

Los sistemas actuales son herramientas que aprenden, no sujetos que deciden. Esa dependencia de objetivos externos convierte la autonomía en un espejismo: lo que percibimos como independencia de la máquina es, a nivel práctico, una delegación de decisiones humanas

a procesos automatizados. Dicha naturaleza delegada de la autonomía tecnológica tiene efectos éticos y políticos relevantes: fomenta la falta de rendición de cuentas y dificulta la supervisión institucional porque cuando el error lo comete un algoritmo, ¿quién responde? Cuando una IA rechaza una solicitud de crédito, de una subvención, diagnostica una enfermedad o prioriza un contenido político determinado en una red social, la decisión se le puede atribuir a la máquina como si esta actuara al margen de sus programadores o de las instituciones que la implementan, lo cual transfiere el poder decisorio a sistemas que no rinden cuentas y reconfigura de este modo la cadena de responsabilidad. Así, cuando las empresas afirman que su modelo *decide* por razones técnicas, la frontera entre decisión humana y resultado propuesto por la máquina se torna borrosa y obstaculiza la atribución de responsabilidad.

2.2. La opacidad algorítmica

Otro elemento relevante en este debate es la opacidad de la IA moderna. Muchos sistemas de aprendizaje automático, especialmente los basados en redes neuronales profundas, son auténticas «cajas negras»: producen resultados que ni sus propios diseñadores pueden explicar con exactitud. Las transformaciones estadísticas internas —millones de conexiones— no son interpretables, lo que dificulta comprender por qué una IA ha tomado una decisión determinada. Esta falta de explicación plantea un serio desafío para la transparencia democrática y la rendición de cuentas.

Pasquale (2015) advierte que la sociedad contemporánea ha pasado de un modelo de decisión visible a un régimen de «sociedad de la caja negra», donde procesos clave —financieros, sanitarios, administrativos— son gestionados por algoritmos inaccesibles. En concreto, el aprendizaje automático (*machine learning*) hace incomprensible el funcionamiento del sistema no solo para la ciudadanía y las personas afectadas en particular, sino también para aquellas con formación específica y expertas en computación (Burrell, 2016, p. 10). Cuando el razonamiento de una máquina es ininteligible, desaparece la posibilidad de evaluar su motivación o de revisar sus decisiones. La tutela judicial efectiva, fundamental en un Estado de derecho, se vacía de contenido si no se puede saber cómo y por qué se ha tomado una decisión automatizada.

Para afrontar este problema, Floridi *et al.* (2018) han propuesto el principio de *explicabilidad* como uno de los cinco pilares éticos de la IA, junto con la beneficencia, la no maleficencia, la autonomía y la justicia. La explicabilidad combina dos exigencias: inteligibilidad (que da respuesta a la pregunta: ¿cómo funciona?) y responsabilidad (que nos permite responder a: ¿quién es responsable de la forma en la que opera el sistema?). Ello requiere transparencia no solo en las operaciones y resultados, sino también en los objetivos iniciales que dirigen a la máquina (Floridi, 2024, p. 316). Sin estos elementos, la IA se vuelve incompatible con las exigencias de transparencia propias de las democracias deliberativas. La explicabilidad no debe

entenderse solo como una obligación técnica, sino como una exigencia política para que la ciudadanía conozca las reglas del juego en una esfera pública cada vez más digitalizada.

2.3. La ilusión de neutralidad

Finalmente, la IA se presenta con frecuencia como una tecnología neutral, guiada por datos y cálculos objetivos. Su eficacia y su creciente integración en la vida social y privada han generado una ilusión de racionalidad técnica. A medida que la IA y los algoritmos se incorporan a nuestra cotidianidad como en los resultados de un motor de búsqueda o en las recomendaciones de una plataforma de vídeo por demanda, o incluso a procesos de decisión en ámbitos como la sanidad, la educación, el empleo o la justicia, tendemos a atribuir a la IA una objetividad que no posee. Esta creencia es una de las más persistentes y, a la vez, una de las más peligrosas. La neutralidad tecnológica es una ilusión. Los algoritmos no son neutrales: están diseñados sobre datos históricos que reflejan las estructuras de poder y los sesgos de las sociedades en que se producen (Noble, 2018).

En efecto, los datos masivos con los que se entrena a la IA no son fragmentos de una realidad objetiva, sino una representación de hechos recogida, filtrada y clasificada por seres humanos con una cosmovisión particular. La IA es más una herramienta cultural que científica y, por tanto, encierra en su arquitectura decisiones sociales y políticas. Los algoritmos aprenden de correlaciones pasadas, y si el pasado está marcado por desigualdad, la IA la reproducirá. Por ejemplo, los sistemas predictivos en el contexto laboral o crediticio pueden perpetuar la discriminación de grupos histórica y estructuralmente excluidos (Barocas y Selbst, 2016).

Esta falsa neutralidad tiene implicaciones políticas. Si se asume que las decisiones de la IA son objetivas, se reduce el espacio para la crítica y se debilita la deliberación pública. Las decisiones automatizadas se vuelven dogmáticas, revestidas de una autoridad técnica que las hace parecer incuestionables. La falta de transparencia que caracteriza a esta nueva «sociedad de la caja negra» gobernada por algoritmos fomenta la sensación de neutralidad y permite a empresas e instituciones ejercitar un poder social y económico opaco, complejo y sin la necesaria rendición de cuentas; se trata de un poder que está más allá del político y que no es neutral (Pasquale, 2015). López de Mántaras y Brunet (2024, p. 20) sintetizan el problema de forma precisa: el problema no está en la IA, sino en quién está detrás de ella. No existe algoritmo sin contexto. Las máquinas aprenden del ser humano, y por tanto, reproducen nuestras jerarquías, prejuicios y aspiraciones. La neutralidad es imposible en tanto que toda selección de datos y todo diseño de modelo implica decisiones sesgadas.

Desmontar la ilusión de neutralidad y de autonomía es esencial para entender las amenazas que la IA plantea a la democracia y a los derechos humanos. Si las decisiones automatizadas reproducen sin cuestionamiento las estructuras de poder, la deliberación pública se vacía de

contenido. El reto no consiste solo en regular los algoritmos, sino en democratizar el poder informacional que los sustenta.

3. Internet y el capitalismo de vigilancia: de la utopía digital a la distopía algorítmica

Internet, en sus orígenes, fue percibido como una promesa de emancipación. El ciberespacio simbolizaba el triunfo de la comunicación horizontal frente al control vertical de los Estados y de los grandes medios. Se habló de una utopía digital en la que la información fluiría libremente, las jerarquías se disolverían y cada individuo podría expresarse y conectar con otros sin mediaciones. Así lo manifestaba Barlow (1996) en su «Declaración de independencia del ciberespacio», en el que describía el nuevo mundo digital como un lugar libertario en el que los gobiernos del mundo industrial y físico carecían de soberanía para regular el comportamiento de los cibernautas.

Esa visión optimista de Internet se apoyaba en una convicción profunda: la tecnología sería intrínsecamente liberadora. El acceso abierto al conocimiento y la interconexión global parecían prometer un futuro más democrático y cooperativo. Sin embargo, apenas dos décadas después, el paisaje digital se transformó en su contrario. El ciberespacio descentralizado dio paso a un ecosistema dominado por unas pocas plataformas que concentran la información, los datos y la atención de miles de millones de usuarios. Lo que en sus inicios fue un espacio de libertad se ha convertido en un entorno de vigilancia, manipulación y control, sostenido por un modelo económico basado en la extracción sistemática de datos personales. Ya en plena ola de ciberoptimismo, Rheingold (1993) advertía de que las comunidades virtuales podían revitalizar la democracia o bien llevarnos hacia las fauces de un oscuro sustituto del discurso deliberativo.

Este cambio de paradigma, que Morozov (2011) denominó «el desengaño de internet», no supuso el fracaso técnico de la red, sino su captura por las lógicas del mercado y del poder político. La promesa emancipadora se transformó en una sofisticada infraestructura de control. Donde antes había anonimato, hoy hay trazabilidad total; donde había diversidad, hoy predominan burbujas de información (*filter bubbles*); donde había descentralización, hoy hay monopolios digitales. Para comprender este giro, es necesario analizar el modelo económico que lo sostiene: el capitalismo de vigilancia, expresión acuñada por Zuboff (2019) para describir la nueva fase del capitalismo informacional.

3.1. Del sueño ciberutópico al sistema de extracción

El internet inicial se construyó sobre ideales libertarios y comunitarios. La cultura *hacker* de los años ochenta y noventa promovía la apertura del conocimiento y el intercambio colaborativo. Los primeros foros, blogs y wikis encarnaban esa filosofía: conocimiento compartido, horizontalidad y ausencia de intermediarios. No se trataba solo de una infraestructura tecnológica, sino de una ideología que asociaba la descentralización digital con la democracia. Morozov (2011) explica cómo esa *ciberutopía* impregnó también el nacimiento de las

plataformas de redes sociales, anunciadas como herramientas propicias para desarrollar sistemas democráticos y, por ende, para acabar con el autoritarismo en el mundo. Hubo quien afirmó que Twitter había sido esencial para movilizar las protestas iraníes de 2009, cuando realmente solo unos pocos miles de tuits fueron publicados desde Irán en esos momentos ya que dicha plataforma no gozaba de popularidad en el país. El autor asocia dicha ciberutopía con un ciberidealismo cargado de ciberinocencia.

En realidad, el ciberespacio fue pronto ocupado por el poder económico y político. El ideal libertario de los cibernautas pronto se convirtió en el libertarismo pragmático de las grandes corporaciones tecnológicas. A medida que internet se convirtió en un espacio central de la vida social, estas empresas comprendieron que los datos personales —navegaciones, preferencias, interacciones— constituían un nuevo recurso económico. Así surgió el modelo de negocio que domina la economía digital contemporánea: la monetización de la información sobre el comportamiento humano.

Zuboff (2019) describe este fenómeno como un proceso de acumulación por vigilancia. Google comenzó almacenando datos de sus usuarios sin saber qué hacer con ellos en un principio. Se trataba de un excedente inutilizado. Posteriormente, decidió emplearlo para mejorar el servicio de su buscador y personalizar la experiencia. Más tarde, dicha información acumulada se destinó a mejorar la rentabilidad de los anuncios, convirtiendo el arte de la publicidad en una ciencia. La compañía pronto observó que servir las necesidades de las personas era menos lucrativo y, por lo tanto, menos importante, que vender predicciones sobre su comportamiento futuro, de forma que el excedente de datos personales, que Zuboff (2019, pp. 92-95) denomina «activo de vigilancia» se ha convertido en una materia prima crítica en la búsqueda de «beneficios de vigilancia» que son finalmente traducidos en «capital de vigilancia». No somos ni los sujetos que crean valor ni tampoco el producto, sino el objeto desde el que se extrae la materia prima para su posterior uso en fábricas predictivas de Google. Así, la explotación de nuestra información convierte nuestra vida en datos de comportamiento para mejorar el control y dominio de otros —los compradores, que pueden ser publicistas o cualquier otro actor interesado— sobre nuestra vida. ¿Por qué querría un buscador como Google invertir en dispositivos como relojes inteligentes, mapas, coches autónomos o sistemas operativos de teléfonos móviles? Precisamente para recopilar más datos privados de los usuarios. Un teléfono barato con un buscador es todo lo que se necesita para obtener la información del mundo (Zuboff, 2019, p. 133).

Pronto se unieron a Google otras grandes compañías como Facebook o Amazon. Estas, junto con redes sociales como X o TikTok, extraen continuamente datos sobre la conducta de los usuarios, que son procesados mediante algoritmos predictivos y los convierten en «productos de predicción» que se venden en mercados de comportamiento futuro. Este modelo supone una ruptura histórica. Si el capitalismo industrial explotaba el trabajo humano, el capitalismo

de vigilancia explota la experiencia humana. Cada clic, cada desplazamiento, cada emoción expresada en línea se convierte en materia prima para la predicción y la manipulación. El espacio digital ya no es un territorio de libertad, sino un laboratorio en el que se estudia, anticipa y orienta la conducta de los individuos.

3.2. El nuevo régimen de vigilancia y control

El capitalismo de vigilancia se sostiene sobre tres pilares: extracción masiva de datos, asimetría informacional y predicción conductual.

En primer lugar, la extracción masiva convierte la vida cotidiana en una fuente inagotable de información. Los dispositivos inteligentes, las redes sociales y los servicios digitales recopilan datos no solo de lo que hacemos, sino también de lo que sentimos, medimos y pensamos. Helbing (2021) advierte que la digitalización ha generado una sociedad en la que cada acción deja huella y puede ser analizada. Esta vigilancia omnipresente se normaliza bajo el pretexto de la personalización o de la comodidad: el usuario acepta ser observado a cambio de eficiencia o entretenimiento. Este mismo autor (2019, p. 53) estima que existen en la actualidad entre 3000 y 5000 archivos personales de casi toda la población en Occidente, información que permite a quien esté en su posesión conocer cómo piensa y siente la persona concreta, a qué partido votará y qué tipo de transacción comercial es más probable que realice en el futuro. Ello coincide con la afirmación de Pariser (2011, p. 43) acerca de que, hace más de una década, una sola compañía —Axiom— tenía información sobre el 96 % de la población estadounidense y de otras quinientas millones de personas en el mundo acerca del nombre de sus familiares, sus direcciones postales actuales y pasadas, cada cuánto pagaban sus facturas, si tenían animales domésticos y su raza, si eran zurdas o diestras, el tipo de medicación que tomaban y un listado de hasta 1500 ítems informativos.

En segundo lugar, la asimetría informacional confiere a las plataformas un poder sin precedentes. Mientras los usuarios apenas conocen cómo se gestionan sus datos, las empresas tecnológicas disponen de herramientas para modelar sus decisiones. Aquí conviene destacar de nuevo la «sociedad de la caja negra» de Pasquale (2015) para describir el entorno donde los algoritmos deciden qué vemos, compramos o creemos, sin transparencia ni posibilidad de réplica. Esa opacidad no solo protege secretos comerciales, sino que perpetúa la subordinación cognitiva de la ciudadanía ante las grandes corporaciones.

Por último, la predicción conductual transforma la vigilancia en manipulación. Zuboff (2019) demuestra que las plataformas no se limitan a observar el comportamiento, sino que intervienen activamente en él. Los algoritmos de recomendación ajustan los contenidos para maximizar el tiempo de atención; los sistemas de publicidad personalizada inducen emociones concretas para incrementar el consumo; los mecanismos de notificación generan dependencia.

Este control no se ejerce mediante la censura directa, sino mediante la gestión invisible de la atención.

En este punto, las advertencias de Cass Sunstein (2017) sobre la fragmentación del espacio público cobran un valor decisivo. En su análisis de las redes sociales, Sunstein observa que la lógica de la personalización crea «cámaras de eco» donde las personas solo se exponen a ideas afines. Pariser (2011) lo denomina *filter bubble*: los algoritmos priorizan los contenidos que confirman nuestras creencias previas y ocultan los que podrían desafiarlas. Así, la deliberación pública —base de la democracia— se sustituye por la reafirmación individual.

La consecuencia política de este fenómeno es una erosión del pluralismo. Si la ciudadanía deja de compartir un espacio informativo común, la posibilidad de un debate racional se desvanece. López de Mántaras (2025) advierte que «sin verdad verificable no hay deliberación posible, y sin deliberación la democracia pierde sentido». Las redes sociales, al privilegiar la emoción sobre la razón y la viralidad sobre la veracidad, debilitan la infraestructura cognitiva sobre la que descansa la legitimidad democrática.

3.3. De la vigilancia al autoritarismo digital

El capitalismo de vigilancia no solo transforma la economía, sino también la política. Las mismas tecnologías que permiten conocer al consumidor facilitan controlar al ciudadano. Es erróneo pensar que internet y las redes sociales son herramientas emancipadoras que amenazan la existencia de los Estados autoritarios en el mundo. De hecho, es todo lo contrario. Ofrecen vías efectivas para afrontar los retos que amenazan su supervivencia (Feldstein, 2019). Morozov (2011) demuestra que el acceso a medios de comunicación, incluso los del enemigo, no fomentan el cambio de régimen, sino todo lo contrario. Pone de ejemplo a la República Democrática Alemana para afirmar que su población estuvo expuesta durante décadas a información occidental, pero lejos de empoderar y liberar a su ciudadanía, la hizo más complaciente y se evidenció una actitud más positiva hacia el propio sistema. La clave estuvo en que su consumo se dirigía a programas de entretenimiento. Esta experiencia ha enseñado a otros Estados autoritarios a emplear las nuevas tecnologías de la información para mantener su poder. Sus gobiernos han descubierto que la mejor forma de marginar literatura disidente no es mediante la censura, lo cual puede suscitar más interés en el producto prohibido, sino en inundar el mercado de libros que entretengan a la ciudadanía. Lo idóneo, por tanto, no es censurar explícitamente, sino introducir en el mercado de las ideas más material para nublar y ocultar así con más información las ideas problemáticas para el poder. Para distraer de la política, lo mejor es el entretenimiento y gracias a las nuevas tecnologías, los Estados autoritarios están pudiendo ganar legitimidad, aplacar las ganas de cambio de las masas y hacerles dedicar su tiempo a ver programas de televisión o escuchar música occidental en lugar de manifestarse por la falta de alimentos o la opresión.

Con la IA, la vigilancia de la población se ha simplificado significativamente. Nunca nadie había dispuesto de tanta información privada de la ciudadanía como sucede hoy con las grandes empresas tecnológicas. Hay datos suficientes para anticipar y contrarrestar la disidencia, para movilizar a personas y colectivos hacia un fin predeterminado e interesado. De la misma manera que se emplea el derecho para el control social, la IA, que se halla al margen del espacio de deliberación democrática, puede regular el comportamiento con asombrosa eficacia sin que exista una verdadera consciencia ciudadana al respecto. Un ejemplo paradigmático de ello es el sistema de crédito social chino, que clasifica a las personas según su comportamiento económico y social, otorgando recompensas o sanciones en función de su «fiabilidad». Este sistema, basado en la interconexión de bases de datos y en la puntuación algorítmica, ilustra la fusión entre control político y tecnología de predicción.

No obstante, la línea que separa a los regímenes autoritarios de las democracias liberales es cada vez más tenue: también en Occidente los gobiernos recurren a sistemas automatizados de vigilancia con fines de seguridad o gestión de riesgos (Zuboff, 2019, p. 112). El capitalismo de vigilancia ha configurado una arquitectura institucional del ciberespacio propicia para la recopilación de datos, pero no para su protección. Las plataformas, al ofrecer servicios gratuitos, construyen un ecosistema en el que el consentimiento se convierte en una verdadera ficción jurídica: los usuarios aceptan términos incomprensibles a cambio de su acceso a herramientas cada día más necesarias para desarrollar actividades cotidianas esenciales como la socialización, el entretenimiento o el trabajo. Dicho consentimiento forzado acaba por legitimar dicha vigilancia permanente.

En Europa, algunos intentos regulatorios, como el Reglamento General de Protección de Datos¹³⁶ o la Ley de IA de la Unión Europea¹³⁷, buscan equilibrar innovación y derechos fundamentales, pero las normas, debido al *tempo* democrático, llegan siempre tarde en su regulación del ciberespacio. Por otra parte, las grandes compañías tecnológicas han crecido al amparo de la desregulación y del libertarismo, creando una ideología *ad hoc* que se sitúa en contra de la acumulación del poder estatal y de su potestad para regular el ámbito privado, es decir, que se oponen a otorgar autoridad a una entidad que representa a la ciudadanía en su conjunto, pero optan por otorgar el poder a corporaciones privadas que no sirven a intereses

136 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 (Reglamento general de protección de datos).

137 Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 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).

mayoritarios, sino al ánimo de lucro de sus propietarios. En este contexto, existe una clara asimetría en el intento de creación de un contrato social digital. Las empresas del sector se caracterizan por su opacidad y por una innovación constante que dificulta cualquier intento de regulación normativa. Además, el legislador carece de la formación necesaria para comprender la amplitud del fenómeno. Asimismo, el poder de las plataformas no reside solo en su capacidad para recopilar datos personales, sino en el hecho de que generan una importante dependencia social y económica al ofrecer productos y servicios cada día más relevantes para la ciudadanía (De Gregorio, 2022). Por si fuera poco, los Estados, en muchos casos, son usuarios de las mismas tecnologías que deberían regular.

El riesgo de este escenario es una deriva hacia lo que Rouvroy y Berns (2013) denominan «gobernanza algorítmica»: un modo de gobierno basado en la predicción y la anticipación, donde las decisiones se adoptan antes de que los problemas ocurran, al estilo *Minority Report*. En este modelo, la libertad deja de ser un principio y se convierte en una variable gestionable por quien conoce al usuario y sabe cómo inducirle a determinados comportamientos. La IA no se limita a observar la realidad: la produce, determinando qué conductas son deseables y cuáles no. Este tipo de control es especialmente peligroso porque se ejerce sin coerción visible. A diferencia de los totalitarismos del siglo XX, que imponían la obediencia mediante el miedo, el autoritarismo digital seduce mediante la comodidad, como en el caso de la República Democrática Alemana destacado por Morozov (2011) y la eficiencia. Nos entrega servicios personalizados, rutas optimizadas, diagnósticos inmediatos y entretenimiento infinito, a cambio de una vigilancia constante. Como señala Han (2014), vivimos en una sociedad del rendimiento y de positividad desmedida en la que las personas se someten voluntariamente a los imperativos de la autoexplotación y del control algorítmico.

3.4. Consecuencias para la democracia y los derechos humanos

El capitalismo de vigilancia redefine la relación entre ciudadanía, información y poder. En un entorno donde los datos sustituyen a los argumentos y los algoritmos median toda interacción, la autonomía individual y la deliberación colectiva se ven gravemente comprometidas. Desde la perspectiva de los derechos humanos, los riesgos son múltiples. En primer lugar, el derecho a la privacidad se ve erosionado por la vigilancia constante. La frontera entre lo público y lo privado se difumina, y el individuo se convierte en un conjunto de datos interpretados por sistemas automáticos. En segundo lugar, la libertad de expresión se ve condicionada por los mecanismos de visibilidad algorítmica: lo que no se muestra no existe, y lo que se prioriza moldea la agenda pública. En tercer lugar, el principio de no discriminación se vulnera cuando los algoritmos reproducen sesgos raciales, de género o socioeconómicos presentes en los datos históricos (O'Neil, 2017; Noble, 2018).

Pero más allá de estos derechos concretos, lo que está en juego es la capacidad democrática de las sociedades para decidir colectivamente su destino. Si la deliberación pública se sustituye

por la manipulación emocional, si las preferencias políticas se diseñan mediante microsegmentación, y si el espacio informativo está gobernado por intereses opacos, la democracia se vacía de sustancia. La gobernanza algorítmica de la que hablan Rouvroy y Berns (2013) corre el riesgo de sustituir el debate ciudadano por una optimización técnica que predetermina el comportamiento individual y colectivo hacia una serie de fines propuestos por quienes poseen la información privada de la ciudadanía. Para explotar esta libertad de recopilar y usar los datos por parte de las grandes compañías, la democracia debe quedarse al margen (Zuboff, 2019, p. 100). Tal y como afirma esta autora, para acabar con la democracia no hace falta un poder totalitario, un sistema que depende del adoctrinamiento y de la coacción abierta, sino que basta con un poder instrumental que sutil e indirectamente emplee información masiva y predicciones de comportamiento para influir en acciones individuales y grupales sin el conocimiento y consciencia de la(s) persona(s) a la(s) cual(es) se dirige. Se trata de la instrumentalización de la conducta para modificar, predecir, monetizar y controlar. Google necesita usuarios distantes, no una ciudadanía. La democracia es prescindible en este nuevo sistema informacional.

4. Democracia y derechos humanos: deliberación y verdad verificable

La democracia descansa sobre una premisa esencial: la confianza en la posibilidad de un debate racional sustentado en hechos compartidos. No hay ciudadanía sin información, ni deliberación sin verdad verificable (López de Mántaras, 2025). La irrupción de la inteligencia artificial y de los algoritmos en el espacio público amenaza precisamente ese presupuesto: al mediar la comunicación y la información mediante sistemas automatizados opacos y guiados por intereses económicos o políticos, se socava la base cognitiva de la democracia.

4.1. La democracia como deliberación

Dahl (1989) entiende la democracia como un sistema de participación efectiva de una ciudadanía que se halla en una situación de igualdad de oportunidades para manifestar sus preferencias y para incluir nuevos temas en los programas de acción, así como para expresar el motivo de su toma de decisión. Asimismo, el autor determina que una democracia requiere de una comunidad política formada por personas que puedan descubrir y convalidar la elección de los asuntos que son sometidos a debate que mejor sirvan a los intereses de la sociedad. Es decir, la democracia se sostiene bajo los pilares de la igualdad política y de la disponibilidad de información veraz. En su concepción de la poliarquía, Dahl subraya que la legitimidad democrática depende no solo de las instituciones formales, sino de las condiciones sociales que hacen posible un ejercicio informado de la ciudadanía. Habermas (2005) ofrece una visión similar a la de Dahl al concebir la democracia como un proceso de formación discursiva de la voluntad, cuya legitimidad está condicionada a la existencia de un espacio público para el intercambio racional de argumentos donde la ciudadanía pueda participar en condiciones de igualdad. Para ello, es esencial que los individuos puedan distinguir entre hechos y opiniones,

entre verdad y falsedad. Habermas destaca que la igualdad de derechos políticos solo puede respetarse si existe simetría en la libertad comunicativa de todos los miembros de la comunidad política y que el discurso racional que legitima la democracia exige iguales derechos en el acceso a la argumentación y una participación simétrica en el proceso deliberativo que asegure tanto la libertad de elección como de inclusión de las mejores informaciones y razones (2005, pp. 193 y 301).

Sin embargo, el ecosistema digital actual erosiona esos presupuestos. La nueva esfera pública plataformizada (Habermas, 2023) se caracteriza por algoritmos que no se orientan a la verdad, sino a la captación de atención. Su lógica no es deliberativa, sino económica. El modelo de negocio de las plataformas requiere una extracción constante de datos de los usuarios y ello motiva que su diseño esté orientado a maximizar su permanencia en línea. Así, se priorizan contenidos que aumentan la interacción —por su capacidad de provocar indignación, miedo o deseo—, por lo que las redes sociales transforman el espacio público en un mercado emocional. Asimismo, la personalización de la experiencia en forma de filtros (*filter bubbles*) que fomenta el entretenimiento y la comodidad realmente obedece a los mismos propósitos del capitalismo de vigilancia, una especie de arquitectura de control que tiene graves consecuencias para la libertad y la democracia (Sunstein, 2017). Y esto es así porque, según este autor, se fragmenta el debate y se reduce la exposición a puntos de vista divergentes, debilitando el pluralismo cognitivo necesario para el juicio político.

En el ciberespacio, por tanto, proliferan mensajes inverificables, desinformación viralizada y una manipulación emocional que consigue su propósito: mantener el máximo tiempo posible a la persona conectada. Dicha saturación informativa paradójicamente reduce el conocimiento. El entretenimiento, sea cual sea su forma, puede fomentar la pasividad política ciudadana que Morozov (2011) describe en el contexto de la República Democrática Alemana al ser expuesta a los medios de comunicación occidentales. Justo en el momento en el que se incrementa exponencialmente la manipulación informativa que genera réditos políticos y económicos, la ciudadanía se vuelve más pasiva y menos preparada para combatirla. La IA amplifica estos problemas al introducir nuevos niveles de automatización en la generación y difusión de contenidos. Los modelos de lenguaje y las redes neuronales generativas producen textos, imágenes y vídeos indistinguibles de los reales, creando un entorno en el que la veracidad se vuelve difícil de determinar. La aparición de *deepfakes* pueden dañar y amenazar la confianza pública en la información visual o textual (Sunstein, 2025). Si todo puede ser fabricado, nada puede ser verificado.

Esta crisis de la verdad tiene consecuencias políticas profundas. Bobbio (1996) ya advertía que la democracia no se define solo por el número de votantes, sino por la calidad de la información y el debate que preceden al voto. La transparencia, la rendición de cuentas y el control ciudadano dependen de que los hechos puedan ser contrastados. Cuando la verdad se

fragmenta en burbujas informativas personalizadas, el principio de publicidad se sustituye por un régimen de opacidades gestionadas algorítmicamente.

4.2. Libertad de expresión y distorsión algorítmica

La libertad de expresión es un pilar fundamental de la democracia deliberativa, al ser la condición que permite la formación de una opinión pública libre y plural. Sin embargo, se erige como uno de los derechos más directamente afectados por la mediación algorítmica. En teoría, internet permite el acceso a la palabra pública: cualquier persona puede expresar sus ideas y alcanzar audiencias globales. Pero, en la práctica, el acceso efectivo a la visibilidad depende de mecanismos de selección automatizada controlados por unas pocas corporaciones tecnológicas. La cuestión central ya no es tanto quién puede hablar, sino quién puede ser escuchado y bajo qué condiciones de visibilidad se ejerce el discurso público.

Las plataformas digitales no son simples intermediarias neutras en este contexto, sino que configuran el espacio discursivo, ya que sus algoritmos de recomendación y priorización toman decisiones sobre moderación, amplificación o supresión de contenidos que determinan qué voces circulan, cuáles se silencian y qué tipo de debate público es posible. Esta lógica de *curaduría algorítmica* en la que el usuario no es responsable de la circulación de noticias e información (Napoli, 2015) transforma la libertad de expresión en una libertad condicionada por criterios comerciales. La visibilidad se otorga no en función de la relevancia democrática o del valor argumentativo, sino del potencial de generar atención y monetización.

La libertad de expresión se ve claramente limitada por los nichos ideológicos creados por las burbujas de filtros (*filter bubbles*) que acaban por convertir los espacios digitales en universos paralelos (Sunstein, 2017). Si bien hemos empleado los filtros personalizados en nuestra propia selección de los medios de comunicación tradicionales para escoger aquellos que fueran acordes a nuestros gustos e intereses, los filtros digitales introducen tres dinámicas novedosas (Pariser, 2011, p. 10). En primer lugar, el usuario se halla solo en su burbuja. La personalización es tal, que no existen dos experiencias iguales. En segundo lugar, esta burbuja es invisible. No sabemos qué piensa Google de nosotros ni por qué ofrece un contenido y no otro. Al no ser el usuario quien determina los criterios de selección para filtrar la información, se fomenta la ilusión de neutralidad y objetividad de la máquina. Por último, nadie elige entrar en la burbuja. Cuando se escoge un canal de televisión, se toma la decisión activa de decidir qué filtro emplear para ver el mundo. Esto no es así en internet. La personalización avasalla al usuario porque incrementa los beneficios de las plataformas y ello hace que cada vez sean más difíciles de evitar.

Todo ello da lugar a cámaras de eco (*eco chambers*) que dividen a la sociedad en espacios digitales estancos y aislados y que fragmentan el debate público. Sunstein *et al.* (2007) realizaron un estudio para conocer los efectos divisorios de los endogrupos con ideologías

afines. Comprobaron que las opiniones individuales previas al experimento en torno a cuestiones políticas sensibles, como la acción positiva, el cambio climático o las uniones civiles, se extremaron después de que los individuos fueran colocados en grupos de su misma ideología (conservadora o progresista). Las opiniones previas se fortalecían ante la presencia e interacción con gente que pensaba igual (Sunstein, 2017, p. 69). Los filtros en internet funcionan de manera similar. Pariser (2011, p. 15) denuncia que son una especie de autopropaganda invisible que nos adoctrina con nuestras propias ideas y amplifica nuestro deseo por cosas que son familiares. Esta personalización expone a los usuarios a contenido autoconfirmatorio. La democracia no es solo una mera cuestión de pluralismo político, sino que necesita que los distintos puntos de vista dialoguen e interactúen en espacios compartidos.

Esta arquitectura de control propiciada por el capitalismo de vigilancia tiene un doble efecto. Por un lado, favorece la polarización: los contenidos extremistas o sensacionalistas, que despiertan emociones fuertes, circulan con mayor rapidez y obtienen más interacciones. Por otro lado, produce una invisibilización selectiva de voces disidentes o minoritarias, que quedan marginadas por no ajustarse a las métricas de popularidad. De este modo, la gobernanza algorítmica configura el espacio público sin rendir cuentas a criterios democráticos. En este sentido, no es necesaria la censura explícita para acabar con opiniones disidentes o para acallar determinados tipos de expresión. Basta con programas los algoritmos para premiar con visibilidad una corriente determinada para que el resto de los argumentos se vean invisibilizados y olvidados.

Sunstein (2017) considera que la censura es uno de los principales obstáculos para la democracia y la libertad, y por ese motivo estima necesario un sistema de libertad de expresión que permita a la ciudadanía exponerse a contenido que no habría escogido de haberlo sabido, en espacios propicios para los encuentros espontáneos, no anticipados y no planificados, puesto que solo así se podrá prevenir la polarización, la fragmentación y el extremismo de quienes hablan solo con personas que piensan igual que ellas en sus cámaras de eco. Para el autor, la democracia necesita una amplia gama de experiencias comunes para facilitar y fomentar la comunicación y el entendimiento entre personas. Esto es muy difícil de lograr en un ciberespacio regido por la lógica mercantilista de extracción de datos personales.

Habermas (2023) ha advertido recientemente sobre la necesidad de reconstruir un espacio digital que hoy está *plataformizado* y que está distorsionando la percepción de la esfera pública no por la existencia de noticias falsas, sino por el hecho de que dicha información fraudulenta no puede identificarse como tal. Esto supone un grave peligro para el proceso deliberativo de la formación de la voluntad pública y, por ello, exige a las plataformas la asunción de responsabilidades cívicas equiparables a las de los medios tradicionales. La garantía de la pluralidad de exposición, la transparencia en los criterios de selección y el mantenimiento de una esfera pública inclusiva que fomente una opinión pública fundada en la deliberación

simétrica de la ciudadanía no es una mera cuestión política para este autor, sino un imperativo constitucional.

Sin embargo, el peligro no radica únicamente en la manipulación intencional —como la desinformación política o las campañas de odio—, sino en la transformación estructural de la comunicación pública. La libertad de expresión deja de ser un derecho de participación en un debate común para convertirse en una experiencia de consumo individualizado. Como subraya Zuboff (2019), en el capitalismo de vigilancia la palabra ya no es un instrumento de deliberación, sino un recurso de extracción: cada interacción comunicativa se traduce en datos que alimentan la maquinaria de predicción y control. Esta dinámica plantea una paradoja: la tecnología que prometía democratizar la comunicación termina reforzando el control sobre lo que se dice y se escucha. La libertad de expresión, desprovista de visibilidad y veracidad, se convierte en un simulacro.

4.3. Privacidad, autonomía y poder informacional

La personalización en el ciberespacio atrae porque facilita la navegación por el mundo digital y convierte la experiencia del usuario en la más interesante posible, pero ese filtro presenta un inconveniente principal: nuestra intimidad se ve vulnerada y nuestros pasos por internet acaban siendo dirigidos por un código en propiedad de grandes corporaciones. Así, la privacidad se ha convertido en una de las condiciones más frágiles —y al mismo tiempo más decisivas— para la vigencia de los derechos humanos en la era digital. No se trata únicamente de proteger datos personales frente al uso indebido por parte de terceros, sino de preservar la autonomía moral y política del individuo frente a sistemas que recopilan, procesan y predicen conductas a gran escala. En palabras de Becker (2019), reducir la privacidad a un problema individual de control sobre la información equivale a ignorar su dimensión estructural: las amenazas contemporáneas provienen de entornos institucionales y tecnológicos que moldean las decisiones y los hábitos de los sujetos, a menudo sin su conocimiento. La privacidad, por tanto, tiene un carácter social e institucional, no solo personal.

El enfoque de Becker subraya que la vigilancia digital —ya sea estatal o corporativa— no se limita a observar lo que hacemos, sino que configura las condiciones bajo las cuales decidimos. Los sistemas de recomendación, la publicidad personalizada y los algoritmos de segmentación política transforman la privacidad en una cuestión de poder: el poder de anticipar, influir o condicionar comportamientos futuros. Este cambio tiene profundas implicaciones democráticas, porque la libertad de juicio depende de la posibilidad de formar opiniones sin interferencias ni presiones invisibles. La erosión de la privacidad equivale, así, a la erosión de la deliberación libre.

Samonek (2021) propone distinguir entre privacidad personal y privacidad política. La primera se refiere al ámbito íntimo del individuo; la segunda, a la capacidad de la ciudadanía de

participar en la vida pública sin estar sujeta a vigilancia o represalias. Para esta autora, esta última se erige en un pilar de la organización social y en condición de legitimidad de las instituciones políticas. En contextos donde las redes sociales, los servicios digitales y las bases de datos gubernamentales generan trazabilidad constante, la privacidad política se ve amenazada: la ciudadanía interioriza la sensación de ser observada y se autocensura. Este fenómeno, conocido en el ámbito del estudio de la libertad de expresión como *chilling effect*, restringe el pluralismo y la crítica, afectando el núcleo mismo de la democracia deliberativa. Sin privacidad política, la esfera pública se convierte en un espacio condicionado por el miedo o la previsión de consecuencias, lo que altera la autenticidad de la participación ciudadana.

A estas aproximaciones se suma la teoría de la integridad contextual formulada por Nissenbaum (2010), que ofrece un marco normativo útil para evaluar los flujos informativos contemporáneos. Esta autora sostiene que la privacidad no puede definirse por la mera ocultación de datos, sino por el respeto a las normas sociales que regulan su circulación: quién comunica qué, a quién, con qué propósito y bajo qué condiciones. Cuando los algoritmos violan estas normas contextuales (por ejemplo, al reutilizar datos de un ámbito personal en uno político o comercial), se rompe la confianza que sostiene la vida social. La integridad contextual, así entendida, no solo protege la intimidad individual, sino la legitimidad del ecosistema informativo sobre el que se asienta la deliberación democrática.

La privacidad, en consecuencia, no es un derecho meramente negativo —el de ser dejado en paz—, sino un derecho positivo que posibilita la autonomía y la autodeterminación informacional. Su vulneración no solo expone al individuo, sino que distorsiona el funcionamiento de la esfera pública y de los mecanismos de representación. En democracias mediadas por plataformas digitales, la ausencia de privacidad implica que la ciudadanía se convierte en un objeto de cálculo y experimentación, donde la frontera entre información y manipulación se vuelve difusa. En el contexto específico del capitalismo de vigilancia denunciado por Zuboff (2019), el sujeto deja de ser autor de sus decisiones para convertirse en objeto de análisis y modelización. Los algoritmos aprenden de los patrones de comportamiento y anticipan las acciones futuras, interviniendo en ellas mediante *nudges* o empujones conductuales. Este tipo de manipulación invisible, descrita también por Sunstein y Thaler (2008), plantea un desafío para la autonomía moral: las decisiones aparentemente libres pueden estar orientadas por estructuras algorítmicas que moldean el entorno de elección.

El deterioro de la privacidad, por tanto, compromete la igualdad política en su raíz. Cuando ciertos actores —corporaciones tecnológicas, agencias estatales o partidos— disponen de información privilegiada para segmentar, persuadir o predecir comportamientos, la relación entre la ciudadanía y el poder deja de ser simétrica. Esta asimetría informacional erosiona la confianza ciudadana, pues la vigilancia permanente socava la creencia de que el espacio público es un ámbito compartido y equitativo. La consecuencia última es la pérdida de lo que Habermas

exigía para la posibilidad de un discurso racional: sin privacidad, la deliberación se transforma en una interacción condicionada por incentivos y miedos, y el ciudadano deja de ser un sujeto autónomo para convertirse en un objeto de perfilado algorítmico. La democracia corre el riesgo de devenir en una tecnocracia donde el protagonismo, lejos de corresponder a una ciudadanía en la que todas las personas puedan tomar decisiones, sobre todo se vincule a solo aquellas expertas que entiendan de los asuntos que se abordan (Bobbio, 1996, p. 27).

Así, el derecho a la privacidad constituye un pilar de la democracia tanto como la libertad de expresión o el derecho a la información. La protección de la privacidad política y de la integridad contextual debe entenderse como un requisito estructural del proceso democrático, no como un lujo individual. En un entorno digital donde la recopilación masiva de datos es el combustible del modelo económico y del control social, la defensa de la privacidad se convierte en defensa de la libertad misma.

4.4. Desigualdad y discriminación algorítmica

La irrupción de la IA y de los sistemas algorítmicos ha transformado radicalmente los mecanismos mediante los cuales se produce, distribuye y ejerce el poder. Si bien estas tecnologías prometen eficiencia, objetividad y neutralidad, múltiples investigaciones han demostrado que los algoritmos pueden reproducir e incluso amplificar desigualdades sociales existentes, generando nuevas formas de discriminación menos visibles, pero igualmente lesivas para los derechos humanos. La cuestión no es únicamente técnica, sino profundamente política: los algoritmos no solo calculan, sino que clasifican y jerarquizan a las personas en función de criterios que reflejan sesgos estructurales inscritos en los datos con los que fueron entrenados.

Una de las investigaciones más relevantes en este contexto es la llevada a cabo por Noble (2018), quien demuestra cómo los motores de búsqueda reproducen estereotipos raciales y de género bajo la apariencia de neutralidad técnica. Esta autora explica que cuando buscamos en Google, no accedemos a una representación neutra del conocimiento, sino a un sistema mediado por intereses económicos y sociales. Los algoritmos priorizan resultados basados en criterios de rentabilidad y popularidad, lo que produce un sesgo sistemático que refuerza desigualdades históricas. Así, la tecnología no solo refleja el mundo social, sino que lo reconfigura, perpetuando estructuras de dominación bajo el ropaje de la objetividad. Se trata, en palabras de dicha autora, de verdaderos «algoritmos de opresión».

Al entrenarse con datos históricos que reproducen los sesgos presentes en la sociedad, se ha evidenciado que los algoritmos utilizados para conceder créditos, seleccionar personal o asignar recursos públicos tienden a penalizar a los grupos históricamente desfavorecidos, perpetuando la desigualdad bajo una apariencia de objetividad técnica (O’Neil, 2017). Eubanks (2018, pp. 11-13) ahonda en la problemática cuando describe cómo la gestión automatizada

de servicios sociales genera exclusiones sistemáticas de quienes más los necesitan. Los sistemas predictivos de IA tienen un impacto destructivo sobre las minorías de escaso poder adquisitivo y sirven para esconder la pobreza de una clase media necesitada de una distancia ética para tomar decisiones inhumanas. En este sentido, los algoritmos no son neutrales; encarnan valores y prioridades sociales. Están sesgados. Si se diseñan sin supervisión ética y sin diversidad de perspectivas, consolidan los privilegios de quienes ya detentan poder económico y tecnológico.

En realidad, la desigualdad está en el mismo origen de la IA. Una pequeña fracción de la humanidad crea y desarrolla tecnologías que transforman por completo la vida del resto (Floridi, 2024). Zuboff (2019, pp. 172 y 182) hace referencia a esta asimetría cuando describe que la división del conocimiento o aprendizaje (*division of learning*) ha reemplazado a la división del trabajo y que aquella implica la exclusión de parte de la población del control de la información. Esta es digital, pero su cantidad excede nuestra capacidad para descifrar su significado. Como solución, se ha planteado el uso de herramientas de IA para crear valor de ese ingente volumen, pero esto solo afianza la posición de privilegio de los capitalistas de vigilancia, aquellos que poseen aquellas herramientas, lo cual solo hace incrementar su poder y moldear la división del conocimiento a su medida.

Según Crawford (2025), nada de esto es un accidente. Esta autora va más allá al afirmar que la IA está diseñada para discriminar y para perpetuar las jerarquías de poder. Además, denuncia que su desarrollo depende de recursos materiales, energéticos y humanos desigualmente distribuidos. La extracción de datos, minerales y mano de obra precaria en el Sur global revela que detrás de la *nube* digital se esconde una infraestructura física profundamente desigual. La discriminación algorítmica, en este sentido, no se limita al sesgo informático: es también un fenómeno ecológico y económico que refleja relaciones de poder a escala planetaria.

Frente a esta realidad, la UNESCO ha propuesto principios de justicia algorítmica que vinculen la ética tecnológica con los derechos humanos. En su *Recomendación sobre la ética de la inteligencia artificial* de 2021 aboga por una IA centrada en la dignidad humana, la diversidad cultural y la equidad, y destaca la necesidad de una gobernanza global que prevenga la concentración de poder tecnológico en manos de pocos actores. El objetivo no es solo evitar el daño, sino redistribuir los beneficios de la automatización y garantizar que la innovación no profundice las brechas existentes.

5. Conclusiones

La inteligencia artificial constituye una de las fuerzas transformadoras más profundas del siglo XXI. Su capacidad para procesar información, automatizar decisiones y modelar comportamientos redefine la relación entre conocimiento, poder y libertad. Sin embargo, como se ha mostrado a lo largo de este trabajo, la IA no es simplemente una herramienta técnica: es

una infraestructura social y política que reorganiza las condiciones mismas de la deliberación democrática y del ejercicio de los derechos humanos. Como señala Floridi (2024, p. 95) corremos el riesgo de que la IA moldee «nuestros entornos físicos y conceptuales».

La distopía más probable no es aquella en la que las máquinas se rebelan y esclavizan al ser humano, sino una mucho más sutil: aquella en la que la IA actúa como el soma de *Un mundo feliz* de Aldous Huxley, un instrumento de consumo, entretenimiento y pasividad que mantiene a la ciudadanía complaciente y manipulada, no por la máquina en sí, sino por quienes la controlan. En este escenario, el ser humano deja de ser sujeto de acción y se convierte en objeto de gestión. Floridi (2024, p. 96) advierte que «el diseño inteligente humano (...) debería desempeñar un papel importante en la configuración de nuestras futuras interacciones con los artefactos inteligentes actuales y futuros y con nuestros entornos compartidos. Al fin y al cabo, hacer que la estupidez trabaje para uno es un signo de inteligencia». La cuestión, entonces, no es si la IA sustituirá al hombre, sino si el ser humano conservará la capacidad de diseñar y gobernar inteligentemente su propio entorno informacional.

La promesa de racionalidad y eficiencia que acompaña a la IA encierra también una paradoja. Aquello que parecía destinado a ampliar nuestras capacidades cognitivas y sociales amenaza con limitar nuestra autonomía moral y colectiva. El problema no reside en la tecnología en sí, sino en el modo en que se ha integrado en el ecosistema digital contemporáneo: un modelo basado en la extracción de datos, la manipulación de la atención y la concentración del poder informacional en manos de unas pocas corporaciones. Zuboff (2019) lo denomina capitalismo de vigilancia, pero más allá del diagnóstico económico, lo que está en juego es un cambio de paradigma político.

La IA, al operar sobre la vida cotidiana y sobre los flujos de información, se convierte en un poder constituyente: no solo regula comportamientos, sino que configura los marcos de interpretación de la realidad. En términos habermasianos, altera las condiciones comunicativas que hacen posible el entendimiento mutuo. Cuando la deliberación se sustituye por la predicción, la política se convierte en administración técnica. López de Mántaras (2025) advierte que sin verdad verificable no hay deliberación posible. Esa afirmación sintetiza el núcleo del desafío actual. La inteligencia artificial amenaza la democracia no porque la suplante con autoritarismo visible, sino porque socava silenciosamente los cimientos epistemológicos sobre los que se erige: la confianza en la posibilidad de un debate público racional. Si la información se fragmenta, si la visibilidad depende de algoritmos opacos, si la verdad se vuelve indistinguible de la simulación, la ciudadanía pierde su capacidad de juicio y, con ella, su soberanía.

Esta pérdida se agrava en el terreno de la libertad de expresión, donde la censura directa ha sido sustituida por un ranking de opiniones. En el capitalismo de vigilancia, las plataformas

jerarquizan los discursos no por su valor racional o democrático, sino por su rentabilidad algorítmica. Aquellas opiniones o movimientos sociales que se alejan de los intereses de las grandes corporaciones tecnológicas son relegados a la irrelevancia digital, aunque sean mayoritarios o socialmente legítimos. Con el tiempo, lo que fue pensamiento común puede devenir minoritario simplemente por el impulso algorítmico de visibilidad que favorece narrativas convenientes al mercado. De este modo, se configura una opinión pública diseñada por agentes privados, una simulación de pluralismo que sustituye la libertad real por un espejismo de participación.

A esta dinámica se añade la erosión de la serendipia, que, según Sunstein (2017), constituye la esencia de una esfera pública viva y democrática. La lógica de la personalización algorítmica elimina el encuentro fortuito con la diferencia, confinando a los ciudadanos en cámaras de eco que refuerzan sus creencias y anulan la deliberación. La democracia, que depende de la exposición a la alteridad, se empobrece cuando la tecnología convierte la diversidad informativa en un producto de consumo a medida.

En el ámbito de la desigualdad, la inteligencia artificial ha amplificado brechas preexistentes. La desigualdad y la discriminación algorítmica constituyen uno de los desafíos éticos más apremiantes de la era digital. La promesa de objetividad tecnológica se ha revelado ilusoria: los algoritmos no operan en el vacío, sino en contextos sociales atravesados por jerarquías de raza, género, clase y territorio. Reconocer este hecho implica abandonar la idea de neutralidad técnica y asumir que la justicia —también la digital— requiere deliberación, transparencia y control democrático. Una sociedad verdaderamente democrática no puede delegar en sistemas opacos decisiones que afectan al destino de las personas. La rendición de cuentas algorítmica no es una cuestión técnica, sino una exigencia política y moral de los derechos humanos en el siglo XXI.

Frente a este panorama, la defensa de los derechos humanos frente a este nuevo escenario exige una doble estrategia: ética y política. Ética, porque implica reconsiderar los valores que deben guiar el diseño y uso de la IA; política, porque requiere instituciones capaces de garantizar que esos valores se traduzcan en normas efectivas. Desde el punto de vista ético, el primer paso consiste en rechazar la ilusión de neutralidad tecnológica. La IA no es un agente imparcial, sino una construcción humana que refleja prioridades sociales y económicas. Floridi *et al.* (2018) propusieron cinco principios para una «IA buena»: beneficencia, no maleficencia, autonomía, justicia y explicabilidad. Estos principios, aunque formulados en términos generales, apuntan a una verdad fundamental: la tecnología debe diseñarse desde la deliberación democrática, no a espaldas de ella.

Desde la perspectiva política, la respuesta más urgente pasa por el control democrático del poder informacional. Esto implica, en primer lugar, transparencia y rendición de cuentas: los

sistemas algorítmicos que afectan derechos deben ser auditables, explicables y sujetos a supervisión pública. En segundo lugar, es necesaria una alfabetización digital y cívica que permita a la ciudadanía comprender cómo operan los sistemas automatizados que median su vida cotidiana. La educación en derechos digitales debe concebirse como una forma contemporánea de ciudadanía activa: solo quien entiende las reglas de la esfera digital puede participar plenamente en ella. Habermas (2022) insiste en que la democracia requiere no solo instituciones inclusivas, sino una cultura política deliberativa capaz de resistir la manipulación informacional. En tercer lugar, resulta imprescindible redistribuir el poder informacional. Las grandes plataformas tecnológicas se han convertido en actores cuasi soberanos que acumulan más datos —y, por tanto, más capacidad de influencia— que la mayoría de los Estados. Esa concentración es incompatible con la idea de igualdad política. Siguiendo a Helbing (2021), la democratización de la IA requiere una constitucionalización del ciberespacio, donde los principios de derechos humanos se apliquen no solo a los Estados, sino también a los agentes privados que configuran la esfera digital global.

Por último, la reconstrucción del espacio público digital exige recuperar la centralidad de la verdad verificable. La veracidad no es una aspiración ingenua, sino un requisito funcional de la democracia. Debe fomentarse un ecosistema informativo basado en la transparencia de las fuentes, la trazabilidad de los contenidos y la responsabilidad de quienes los difunden. La IA puede contribuir positivamente a este fin (por ejemplo, mediante herramientas de verificación o detección de desinformación) si se orienta a reforzar la deliberación y no a explotar la atención.

La crisis actual no es solo tecnológica, sino antropológica. La IA refleja y amplifica nuestras propias contradicciones: el deseo de conocimiento y control, la búsqueda de eficiencia frente a la necesidad de sentido. La respuesta, por tanto, no puede reducirse a la regulación técnica; debe implicar una reflexión sobre el tipo de humanidad y de convivencia política que queremos preservar.

La historia de internet demuestra que ninguna tecnología es neutral ni irreversible. Lo que comenzó como una herramienta de liberación se ha convertido en un dispositivo de vigilancia; pero también puede transformarse de nuevo en un espacio de cooperación y deliberación, si los valores democráticos orientan su desarrollo.

En última instancia, el futuro de la IA no depende de su potencia computacional, sino de la sabiduría política con la que sepamos gobernarla. Restablecer la deliberación y la verdad verificable en la era algorítmica implica recuperar la confianza en la razón pública, en la transparencia institucional y en la dignidad de cada ser humano como sujeto de decisión, no como objeto de cálculo. Esa es, quizás, la tarea más urgente de nuestro tiempo: evitar que la

IA se convierta en una distopía para los derechos humanos y hacer de ella una herramienta al servicio de la libertad y la justicia.

Bibliografía

Barlow, J. P. *A Declaration of the Independence of Cyberspace*. Electronic Frontier Foundation, 1996. Disponible en: <https://www.eff.org/cyberspace-independence>

Becker, M. «Privacy in the digital age: Comparing and contrasting individual versus social approaches towards privacy». *Ethics and Information Technology*. 2019, vol. 21, núm. 4, 307-317. DOI: <https://doi.org/10.1007/s10676-019-09508-z>.

Bobbio, N. *El futuro de la democracia*. México D. F.: Fondo de Cultura Económica, 1996.

Burrell, J. «How the machine ‘thinks’: Understanding opacity in machine learning algorithms». *Big Data & Society*. 2016, vol. 3, núm. 1, 1-12. DOI: <https://doi.org/10.1177/2053951715622512>

Crawford, K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press, 2021.

Dahl, R. A. *La democracia y sus críticos*. Barcelona: Paidós, 1989.

De Gregorio, G. *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society*. Cambridge: Cambridge University Press, 2022.

Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin’s Press, 2018.

Feldstein, S. «Artificial Intelligence and Digital Repression: Global Challenges to Governance». *Conference Paper*. International Studies Association Annual Meeting, marzo de 2019.

Floridi, L. *The 4th Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press, 2014.

Floridi, L. et al. «AI4People — An Ethical Framework for a Good AI Society». *Minds and Machines*. 2018, vol. 28, núm. 4, 689-707. DOI: <https://doi.org/10.1007/s11023-018-9482-5>.

Floridi, L. *Ética de la inteligencia artificial*. Barcelona: Herder, 2024.

Han, B.-C. *La sociedad del cansancio*. Barcelona: Herder, 2014.

Habermas, J. *Facticidad y validez*. Madrid: Trotta, 2005.

Habermas, J. *A New Structural Transformation of the Public Sphere and Deliberative Politics*. Cambridge: Polity Press, 2023.

Helbing, D. *Towards Digital Enlightenment. Essays on the Dark and Light Sides of the Digital Revolution*. Cham: Springer, 2019.

Helbing, D. *Next Civilization: Why We Need a Global Constitution*. Cham: Springer, 2021.

López de Mántaras, R. «Intel·ligència artificial: De la màgia a la realitat». *Revista de Tecnologia*. 2022, núm. 10, 4-9. DOI: <https://doi.org/10.2436/20.2004.01.32>

López de Mántaras, R. y Brunet, P. «¿Qué es la inteligencia artificial?». *Papeles de relaciones ecosociales y cambio global*. 2024, núm. 164, 13-21.

López de Mántaras, R. «Cómo la IA está minando las democracias y cómo protegerlas mejor». *La Vanguardia*, 7 julio 2025. Disponible en: <https://www.lavanguardia.com/ciencia/20250707/10857902/como-ia-esta-minando-democracias-como-protegerlas-mejor.html>

Morozov, E. *The Net Delusion: The Dark Side of Internet Freedom*. New York: PublicAffairs, 2011.

Napoli, P. «Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers». *Telecommunications Policy*. 2015, vol. 39, núm. 9, 751-760. DOI: <https://dx.doi.org/10.2139/ssrn.2481886>.

Nissenbaum, H. *Privacy in Context. Technology, Policy, and the Integrity of Social Life*. Stanford: Stanford University Press, 2010.

Noble, S. U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press, 2018.

O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Penguin, 2017.

Pariser, E. *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin, 2011.

Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press, 2015.

Rheingold, H. *The Virtual Community: Homesteading on the Electronic Frontier*. Reading, MA: Addison-Wesley, 1993.

Rouvroy, A. y Berns, T. «Algorithmic Governmentality and Prospects of Emancipation: Disparateness as a Precondition for Individuation Through Relationships?». *Réseaux*. 2013, núm. 177, vol. 1, 163-196. DOI: <https://doi.org/10.3917/res.177.0163>.

Samonek, T. «How Is Political Privacy Different from Personal Privacy? An Argument from Democratic Governance». *Diametros*. 2021, vol. 18, núm. 71, 64-77. DOI: <https://doi.org/10.33392/diam.1544>.

Sunstein, C. R. *#Republic: Divided Democracy in the Age of Social Media*. Princeton: Princeton University Press, 2017.

Sunstein, C. R. *Manipulation. What It Is, Why It's Bad, What To Do About It*. Cambridge: Cambridge University Press, 2025.

Sunstein, C. R., Hastie, R. y Schkade, D. «What Happened on Deliberation Day». *California Law Review*. 2007, vol. 95, 915-940.

Sunstein, C. R. y Thaler, R. H. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press, 2008.

Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs, 2019.

Mireia Molina Sánchez. Mecanismos de control para prevenir los sesgos de la herramienta Viogen.

Mireia Molina Sánchez

Ayudante doctora. Universitat de València

[ACCEDER AL VIDEO](#)

Resumen

El presente trabajo¹³⁸ analiza el uso del sistema VioGén, una herramienta desarrollada por el Ministerio del Interior de España para evaluar el riesgo de revictimización en mujeres víctimas de violencia de género. Aunque actualmente no puede considerarse inteligencia artificial según el Reglamento de IA de la Unión Europea, sí se clasifica como un sistema algorítmico automatizado de la administración pública. Si bien VioGén ha demostrado ser eficaz en un elevado porcentaje de los casos, existen situaciones en las que reproduce sesgos o infravalora el riesgo de determinadas víctimas.

Es por ello por lo que, la propuesta central del artículo es considerar la diligencia emitida por VioGén —que contiene el nivel de riesgo asignado y las medidas de protección correspondientes— como un acto administrativo. Esta calificación permitiría que tanto la víctima como el agresor pudieran interponer recursos administrativos y, en su caso, contencioso-administrativos, facilitando así la revisión y corrección de posibles errores del sistema antes de que se produzcan daños irreparables.

De manera que, en lugar de esperar a que la responsabilidad patrimonial del Estado sea reconocida tras una nueva agresión, proponemos dotar al sistema de mecanismos preventivos que refuercen la protección de las víctimas. En este sentido, la posibilidad de impugnación legal de las decisiones automatizadas representa un avance significativo en la garantía de los derechos y la seguridad de las víctimas, así como en la mejora del propio sistema.

¹³⁸ El presente trabajo se enmarca en el proyecto de investigación “la planificación administrativa, ref. PID2023-152781NB-I00” que se encuadra en el Plan Nacional I+D+i: PLANADM 2024-2026, en el proyecto “Algorithmic Law”, ref. CIPROM/2024/16, que se enmarca en el programa Prometeu Generalitat Valenciana: AILAW, 2025-2026 y del convenio C036/23-OT entre la entidad pública empresarial Red.es agrupación de entidades formada por la Universidad del País Vasco/EUSKAL HERRIKO UNIBERSITATEA (UPH/EHU), la Universitat de València, la Fundación Instituto de Investigación Sanitaria “Marqués de Valdecilla”, la Entidad de Gestión de Derechos de los Productores Audiovisuales LA (EGEDA), la Universitat Rovira i Virgili, y la Fundación Hay Derecho, para impulsar la implementación de la carta de derechos digitales en el ámbito de entornos específicos.

Introducción

Dentro de los innumerables ámbitos en los que se utiliza la inteligencia artificial resulta destacable su aplicabilidad en la lucha contra la desigualdad, permitiendo una respuesta rápida y eficaz contra los problemas sociales. No obstante, el empleo de herramientas automatizadas puede ocasionar sesgos y discriminación, produciendo un efecto totalmente contrario al deseado. Esto nos hace poner el foco en la necesidad de desarrollar, no solo sistemas algorítmicos o inteligencia artificial que permita luchar contra las distintas lacras sociales, sino también la implementación de mecanismos que permitan detectar y solucionar posibles sesgos ocasionados por las propias herramientas, lo cual aumenta la discriminación sufrida por estos colectivos que con anterioridad a la aplicación de la herramienta ya sufrían algún tipo de discriminación. Un ejemplo de este tipo de herramientas es VioGén, la cual es empleada en España en la lucha contra la violencia de género. Las mujeres víctimas de violencia de género a las que se le aplica la herramienta VioGén sufren discriminación a manos de su pareja o expareja, pero además pueden ser infra protegidas por la propia herramienta por distintas causas. En este sentido, proponemos que la diligencia que emite VioGén sea considerada como acto administrativo, ya que de esta manera estará sometida al sistema de recursos tradicionalmente previsto para ello. De manera que las mujeres que consideren que la herramienta no les está asignando el nivel de riesgo correcto, podrán recurrir ante el superior jerárquico, a través del recurso administrativo, la incorrección del acto, así como, en caso de no obtener una respuesta que favorable, recurrir el acto administrativo ante la jurisdicción contencioso-administrativa.

¿Qué es VioGén?

VioGén es una herramienta desarrollada por el Gobierno de España en el año 2007, a través de la Instrucción 10/2007, de la Secretaría de Estado de Seguridad, por la que se aprueba el protocolo para la valoración policial del riesgo de violencia contra la mujer en los supuestos de la Ley Orgánica 1/2004, de 28 de diciembre, y su comunicación a los órganos judiciales y al Ministerio Fiscal, en adelante Instrucción 10/2007 (GONZÁLEZ ÁLVAREZ, LÓPEZ OSSORIO y MUÑOZ RIVAS, 2018, pp. 10–11), con el objetivo de detectar el nivel de riesgo que tiene una mujer de volver a ser revictimizada por su pareja o expareja y, en consecuencia, aplicar unas determinadas medidas de protección.

Antes de profundizar en el análisis de la herramienta resulta esencial conocer ante qué tipo de herramienta nos encontramos ya qué en función de ello se le aplicará una normativa u otra. Actualmente, con la entrada en vigor del Reglamento de Inteligencia Artificial, en adelante Reglamento de IA – UE, podemos conocer qué se entiende por inteligencia artificial y, decidir si un determinado sistema debe cumplir con las exigencias contenidas en dicho reglamento –

Hasta el momento no contábamos con ninguna norma nacional ni internacional que definiera la inteligencia artificial ni los sistemas algorítmicos. Únicamente contábamos con definiciones doctrinales como la ofrecida por (SIMÓN CASTELLANO, 2021, p. 94)–. En atención al Reglamento de IA – UE VioGén no podrá ser considerado como inteligencia artificial, ya que se considera una herramienta de menor complejidad al no cumplir con algunos de los requisitos establecidos en las directrices emitidas por la Comisión Europea. Los requisitos que no reúne la herramienta para ser considerada como inteligencia artificial son: la autonomía y la capacidad para inferir resultados (COMISIÓN EUROPEA, 2025, pp. 4–5). No obstante, esto no significa que en pocos meses no podamos hablar de VioGén como un sistema de inteligencia artificial, ya que el Ministerio del Interior anunció su intención de introducir mejoras que conllevarían la consideración de VioGén como inteligencia artificial en atención al Reglamento de IA – UE (LA MONCLOA, 2020).

De manera que, como no podemos decir que hoy en día VioGén, cumpla con los requisitos establecidos por la comisión europea para ser inteligencia artificial, hablaremos de la herramienta como un sistema algorítmico. Ya que un algoritmo puede ser definido como “un procedimiento computacional bien definido que parte de un estado inicial y un valor o un conjunto de valores de entrada, a los cuales se les aplica una secuencia de pasos computacionales finitos, produciendo una salida o solución. Se puede considerar al algoritmo como una herramienta para resolver un cálculo computacional bien especificado” (CORMEN, LEISERSON, RIVEST y STEIN, 2009, p. 26). VioGén encaja a la perfección en dicha descripción, ya que, como hemos avanzado, se encarga de realizar predicciones futuras, empleando para ello una serie de operaciones matemáticas y los datos introducidos.

A nivel nacional la normativa no hace referencia ni a los sistemas algorítmicos ni a la inteligencia artificial, únicamente menciona los sistemas automatizados de las administraciones públicas, lo cual se encuentra previsto en el artículo 41 de la Ley 40/2015, de 1 de octubre, de Régimen Jurídico del Sector Público, en adelante LRJSP. En atención a dicho artículo por actuación administrativa automatizada entendemos “cualquier acto o actuación realizada íntegramente a través de medios electrónicos por una Administración Pública en el marco de un procedimiento administrativo y en la que no haya intervenido de forma directa un empleado público”. En el caso que nos ocupa, VioGén debe ser entendida como un sistema automatizado de la Administración Pública, ya que es la propia herramienta la que toma la decisión acerca del nivel de riesgo de la víctima, así como de las medidas protectoras que deben ser aplicadas en cada uno de los casos – Es importante mencionar que el funcionario interviene para introducir los datos, no obstante, esto es considerado como una intervención indirecta. A mayores, como veremos en el epígrafe siguiente, la herramienta cuenta con un mecanismo de control y es que los agente pueden modificar al alza el nivel de riesgo al alza, cosa que únicamente ocurre en un 5% de las ocasiones. No obstante, en nuestra opinión, la introducción

de este mecanismo de control no es suficiente para entender que no nos encontramos ante un sistema automatizado – SAN MARTÍN SEGURA, 2023, pp. 19–20; ZURITA BAYONA, 2014, pp. 774–778).

De manera que, en atención a la información que tenemos disponible hoy en día sobre VioGén, y la normativa vigente tanto nacional como internacional, podemos decir que VioGén es un sistema automatizado de la Administración Pública, así como que entra dentro de la definición doctrinal de sistema algorítmico.

1.1 Estructura básica de la herramienta

En España, a diferencia de lo que ocurre en otros países en los que existen herramientas parecidas, contamos con dos formularios que son empleados en función de la situación concreta que se denuncia. Cuando la víctima acude a la policía por vez primera los agentes de policía harán uso del formulario VPR 5.0-H (Valoración Policial del Riesgo, versión 5.0-H). Con este protocolo lo que se logra es establecer el nivel inicial de riesgo, así como asignar las medidas protectoras que se consideren oportunas para dicho nivel de riesgo. En los subsiguientes contactos de las víctimas con los agentes se hará uso del formulario VPER4.1 (Valoración Policial de la Evolución del Riesgo, versión 4.1) – Es importante destacar que ambos protocolos fueron introducidos por la Instrucción 10/2007, no obstante, han ido sufriendo modificaciones y adaptándose a los nuevos descubrimientos en torno a la violencia de género, así como también se han ido adaptando a las nuevas formas de violencia. De manera que los cuestionarios han sido modificados a lo largo de los años –. Hasta junio del año 2025 el VPER se componía de dos formularios: la VPER-S, también conocida como VPER sin incidencia y la VPER-C, también conocida como VPER con incidente.

El primero de ellos, la VPER-S, se realiza de forma sistemática cuando transcurre un determinado periodo de tiempo, el cual viene determinado por el nivel de riesgo inicialmente asignado por la herramienta – Los períodos de tiempo van desde las 72 horas en los casos de riesgo extremo, un máximo de 7 días en los casos de riesgo alto, 30 días en los casos de riesgo medio, cada 60 días en caso de riesgo bajo y 90 días en casos de riesgo no apreciado– (SECRETARÍA DE ESTADO DE SEGURIDAD, 2019a, p. 13; LÓPEZ OSSORIO, 2023), y su objetivo es identificar el avance del caso y poder detectar si se produce una disminución del nivel de riesgo y, por tanto, se puede reducir el nivel de riesgo asignado y las medidas protectoras aparejadas, o, si por el contrario, se han dado situaciones que pronostican un aumento del nivel de riesgo y, por tanto, de las medidas protectoras (GONZÁLEZ ÁLVAREZ et al., 2018, p. 63).

La VPER-C, por su parte, es empleada cuando los agentes de policía tienen conocimiento de un nuevo incidente, lo cual puede ser desde una nueva agresión hasta un quebrantamiento de condena. Este cuestionario también es empleado en aquellos casos en que los agentes descubren nuevos hechos relevantes de los que no se tuvo conocimiento en el momento inicial

en el que se empleó la VPR 5.0-H. Finalmente destacar que es posible que sea el juez o fiscal quien inste que se rellene este formulario a la vista de la información que posean sobre el caso en cuestión (GONZÁLEZ ÁLVAREZ et al., 2018, pp. 52–53; ESTÉVEZ MENDOZA, 2021, p. 134).

Desde junio de 2025, con el surgimiento de VioGén 2, la VPER-C ha desaparecido. De manera que únicamente contamos con la VPR 5.0-H y la VPER, que equivale a la anterior VPER-S, esto se debe a que se ha optado por utilizar, en aquellos casos en que se vuelve a producir un incidente de los citados anteriormente para los casos en que se empleaba la VPER-C, la VPR 5.0-H. Esto se debe principalmente a que ambos cuestionarios eran idénticos por lo que se ha tratado de sistematizar un poco más la herramienta (SECRETARÍA DE ESTADO DE SEGURIDAD, 2025).

Una vez que sabemos de qué cuestionarios se compone el sistema debemos conocer su funcionamiento, es decir, cómo son empleados para que arrojen el nivel de riesgo y las medidas protectoras a implementar. Pese a que hablamos de formulario o de cuestionario es importante destacar que en ningún caso son cumplimentados por la víctima, ni tanto solo se les tienen que hacer las preguntas tal y como aparecen en el mismo. Sino que el agente al entrevistarse con la víctima cuando esta va a interponer la denuncia deberá obtener toda la información posible, para posteriormente poder responder el formulario. Para ello deberán realizar preguntas abiertas y que no condicionen las posibles respuestas de la víctima (SECRETARÍA DE ESTADO DE SEGURIDAD, 2019b, p. 5). Cabe destacar que, pese a que la principal fuente de información es el testimonio de la víctima, los agentes no deben conformarse con dicha información, sino que deben consultar muchas otras fuentes, como, por ejemplo: testimonio del agresor, de los familiares, del entorno de la víctima y del agresor, informes médicos, antecedentes policiales y judiciales tanto en España como en el extranjero, etc. (SECRETARÍA DE ESTADO DE SEGURIDAD, 2019a, p. 9). Por lo que lejos de lo que se puede pensar no se trata de una herramienta sesgada que únicamente tiene en consideración la percepción de la víctima, sino que los agentes deben hacer un análisis exhaustivo del caso en el que se valoren todas las fuentes de información posibles.

Todos los datos de una determinada víctima respecto de un agresor concreto se organizan en casos, de manera que lo primero que debe realizar el agente es buscar si ya existe en el sistema un caso para esa víctima y agresor concreto. De no ser así, se dará de alta como un nuevo caso. Mientras que si sí que existe un caso ya registrado se procederá a su actualización (SECRETARÍA DE ESTADO DE SEGURIDAD, 2019a, p. 8).

Finalmente hay que destacar que, pese a que lo largo del texto hemos hablado de que VioGén tiene como objetivo detectar el nivel de riesgo de que una víctima sufra una nueva agresión a manos del mismo atacante, esta no es la única función de la herramienta. Si bien es cierto que esta fue la función inicial con la que se creó la herramienta, las subsiguientes Instrucciones que

han ido modificado la inicial Instrucción 10/2007, han introducido nuevas funcionalidades de la herramienta, que complementan la mera detección del nivel de riesgo. De entre estas funciones podemos destacar: la posibilidad de detectar un nuevo incidente grave o letal (LÓPEZ OSSORIO, MUÑOZ VICENTE, ANDRÉS PUEYO y PASTOR BRAVO, 2020, pp. 25-26), que los menores se encuentren en una situación de riesgo o de vulnerabilidad (SECRETARÍA DE ESTADO DE SEGURIDAD, 2019b), la adopción de medidas especiales de protección en aquellos casos en que no contemos con denuncia de la víctima (SECRETARÍA DE ESTADO DE SEGURIDAD, 2021a), medidas especiales respecto de los agresores persistentes (SECRETARÍA DE ESTADO DE SEGURIDAD, 2021b, p. 4; SECRETARÍA DE ESTADO DE SEGURIDAD, 2023, pp. 4–7) y medidas especiales para los casos resistentes (SECRETARÍA DE ESTADO DE SEGURIDAD, 2022; SECRETARÍA DE ESTADO DE SEGURIDAD, 2021a; SECRETARÍA DE ESTADO DE SEGURIDAD, 2021b). Es importante mencionar que la Instrucción 1/2025 ha derogado todas estas Instrucciones, no obstante, el contenido de ellas aparece sintetizado a lo largo de dicha Instrucción. Por lo que podemos decir que VioGén 2 sigue cumpliendo con todos esos objetivos, que van más allá de la mera detección del nivel de riesgo de la víctima.

1.2 La diligencia emitida por la herramienta

La diligencia que emite VioGén la realiza la propia herramienta sin intervención del funcionario que instruye el caso y se compone del nivel de riesgo asignado por la herramienta, si se da algunas de las situaciones anteriormente mencionadas y de las medidas de protección que deben ser aplicadas. De manera que, bajo nuestro parecer cumple con los requisitos establecidos por la doctrina y la normativa para poder ser considerado como acto administrativo. Ya que, en palabras de García de Enterría, un acto administrativo es una “la declaración de voluntad, de juicio, de conocimiento o de deseo realizada por la Administración en ejercicio de una potestad administrativa distinta de la potestad reglamentaria” (GARCÍA DE ENTERRÍA Y MARTÍNEZ-CARANDE y FERNÁNDEZ RODRÍGUEZ, 2022). Los requisitos establecidos por la Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas, en adelante LPAC, también los cumple ya que en primer lugar se exige que sea emitido por un órgano competente, cosa que se cumple ya que el artículo 31 de la Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género, en adelante LOMPIVG, establece que las FFCCSS serán las encargadas de implementar las medida necesarias para la prevención y protección frente a la violencia de género. En segundo lugar, cumple con la exigencia de motivación, ya que la diligencia emitida por VioGén también incorpora la respuesta que el agente otorgo a cada uno de los ítems que componen el cuestionario, por lo que se puede comprobar si la respuesta del sistema es ajustada. Así mismo, si el agente decide aumentar el nivel de riesgo también aparecerá tal decisión, así como la justificación que le ha llevado a aumentar el nivel de riesgo. En tercer lugar, el requisito de forma también se cumple ya que la diligencia se emite por escrito y en ella

aparece la motivación antes citada. Finalmente, el contenido de la diligencia se adecúa a la normativa sobre violencia de género y al objetivo perseguido por la herramienta.

La relevancia y el gran avance que supondría la consideración de dicha diligencia como acto administrativo reside en que los ciudadanos afectados por la diligencia, en este caso víctima y agresor podrán interponer recurso contra tal diligencia si consideran que no se adecúa a la realidad o si las medidas de protección adoptadas por los agentes no se corresponden con las establecidas en la propia diligencia. Para ello, como veremos con más detalle en el siguiente epígrafe, podrán interponer, en primer lugar, recurso administrativo y, en segundo lugar, recurso contencioso-administrativo. Lo cual permitirá no solo que los afectados manifiesten su desacuerdo con la diligencia emitida, sino también que la Administración pueda subsanar posibles errores cometidos por la herramienta y, en caso de que estemos hablando de una infravaloración del nivel de riesgo, aumentar dicho nivel de riesgo y reducir las probabilidades de que la víctima vuelva a ser revictimizada.

1 Posibilidades de recurso

Como ya hemos avanzado, el hecho de considerar la diligencia emitida por VioGén como acto administrativo tiene como principal ventaja el hecho de que se pueda interponer recurso y subsanar posibles errores y sesgos que contiene la herramienta. Esto es debido a que mediante el recurso se pone de manifiesto un error dando la posibilidad al superior jerárquico o incluso a instancias judiciales de que puedan controlar la adecuación del resultado arrojado por la herramienta.

Antes de entrar a analizar los tipos de recurso que se pueden interponer es importante destacar que en esta ocasión por persona interesada debemos entender tanto a la víctima como al agresor, debido a que ambos tienen un interés legítimo en el resultado que arroja la herramienta. No obstante, ambos harán servir medios de prueba diferentes, debido a que en un tema tan sensible como el que nos ocupa se debe primar la seguridad y la protección de las víctimas. De manera que, mientras que la víctima deberá tener acceso al informe completo emitido por la herramienta, en el que se reflejen todos los ítems tenidos en consideración – Esto es totalmente viable ya que así lo prevé el Derecho Administrativo y ninguna excepción existe cuando se trata de cuestiones penales – (BLASCO DÍAZ, 2010; MARTÍNEZ GARAY et al., 2024).

En atención al artículo 22 de la Ley Orgánica 7/2021, el agresor únicamente deberá tener acceso a aquellos ítems del formulario que tengan relación con él mismo. De manera que el agresor no podrá tener acceso a datos como la existencia de red de apoyo de la víctima, si esta sufre alguna patología psicológica o incluso los datos relativos al domicilio de la víctima. Esto en ningún caso podemos considerar que atente contra el derecho de defensa del agresor, pues

son datos relativos a la víctima que no podrán ser refutados por el agresor y que, en cambio sí que pueden ser empleados por él para revictimizar a la víctima (MARTÍNEZ GARAY et al., 2024).

Recurso administrativo

En el caso que nos ocupa se deberá interponer recurso de alzada, ya que quien suscribe la diligencia que emite VioGén es el agente instructor del caso, por lo que existe un superior jerárquico. De manera que, en caso de que tanto la víctima como el agresor aprecien una incongruencia entre el resultado arrojado por la herramienta y los hechos vividos, podrán interponer este recurso aportando los medios de prueba que tengan en su poder, con el objetivo de que el superior jerárquico revise la valoración efectuada y dictamine si se ha producido un error en la valoración del riesgo. En caso afirmativo, el superior jerárquico podrá ordenar que se efectúe una nueva valoración y, de esta manera, solventar el error contenido en la valoración inicial. De darse, esta situación tendría grandes repercusiones, ya que, al realizarse una nueva valoración, en este caso correcta, si la víctima sufriera una nueva agresión no se le podría imputar a la Administración, ya que esta había hecho todo lo posible para evitarla. En caso contrario, si el superior jerárquico considera que la valoración inicial es correcta, mantendrá la existente y el/la recurrente podrá acudir a la vía jurisdiccional e interponer recurso contencioso-administrativo. De darse una nueva agresión y, en este momento, quedar patente que sí que existía alguna incongruencia en la valoración inicial, se podrá interponer reclamación por responsabilidad patrimonial del superior jerárquico. Cabe destacar que de darse una nueva agresión y no haber interpuesto previamente recurso, se podrá igualmente reclamar la responsabilidad patrimonial de la Administración por los daños ocasionados, siendo la administración recurrida aquella a la que se encuentre suscrita el agente instructor del caso.

Recurso contencioso administrativo

Una vez interpuesto el recurso administrativo, el cual es de carácter obligatorio antes de acudir a la vía jurisdiccional, si el resultado obtenido en la resolución en dicho recurso no es favorable para el/la recurrente, este podrá optar por acudir a la vía contencioso-administrativa. En este momento, podrá solicitar al juez, con base en las pruebas aportadas, que se realice una nueva valoración, pudiendo este decidir si efectivamente hay un error en la valoración inicialmente realizada y, por tanto, procede ordenar que se realice una nueva valoración del riesgo o, si en cambio, considera que no existe ninguna incongruencia y, por tanto, mantiene la valoración inicialmente efectuada. De darse este último caso, y si la víctima sufriera una nueva agresión, tras la cual quedara demostrado que cuando se interpuso recurso la valoración no era correcta porque era inferior al verdadero nivel de riesgo en el que se encontraba la víctima, esta, o sus familiares en caso de fallecimiento, podrán solicitar la indemnización, pero en este caso ya no por responsabilidad patrimonial del Estado, sino por error judicial.

Conclusión

La lucha contra la desigualdad es una ardua tarea que se está llevando a cabo desde el siglo pasado, no obstante, con la llegada de las nuevas tecnologías se le ha dado un nuevo enfoque, ya que estas se caracterizan, teóricamente, por su objetividad y transparencia. Digo teóricamente porque numerosos estudios y noticias han puesto de manifiesto la existencia de conductas discriminatorias en procedimientos en los que no ha intervenido un ser humano sino una máquina (MARTÍNEZ GARAY et al., 2024; SORIANO ARANZ, 2021; MOLINA SÁNCHEZ, 2023). Esto nos lleva a afirmar que hoy en día no contamos con sistemas perfectos ajenos a cualquier sesgo, por lo que resulta indispensable el establecimiento de mecanismos de control que permita identificar aquellos casos en que la herramienta está dando respuestas sesgadas o discriminatorias.

Como hemos avanzado, esto ha ocurrido en numerosos sectores, entre los que podemos destacar la empleabilidad, la concesión de ayudas e incluso en la protección que se otorga a las víctimas de violencia de género. En el presente trabajo hemos decidido centrarnos en el relativo a la protección de las víctimas de violencia de género porque consideramos que es un asunto complejo que requiere especial atención, ya que estamos hablando de mujeres que en primer lugar han sido discriminadas por su pareja o expareja por el mero hecho de ser mujer y, posteriormente, la herramienta empleada para establecer los mecanismos de protección la vuelve a discriminar al no proporcionarle la protección adecuada. Pese a esto, no podemos obviar que se trata de un sistema que, en términos generales, funciona correctamente, ya que según un estudio elaborado en 2019 es capaz de detectar un 85% de los casos en que la mujer se encuentra en una situación de riesgo y efectivamente se produce algún tipo de reincidencia (LÓPEZ OSSORIO, GONZÁLEZ ÁLVAREZ y ANDRÉS PUEYO, 2016).

De manera que, pese a que esta herramienta fomenta en gran medida la seguridad de las víctimas de violencia de género, hemos de ser conscientes de que también comete errores y es aquí donde hay que poner el foco. Hasta el momento los estudios realizados sobre la eficacia de la herramienta se han orientado en mejorarla para que ofrezca mejores resultados. Empero, bajo nuestra opinión, en este tipo de herramientas probabilísticas la posibilidad de que cometan fallos y, resulte discriminatoria con determinados colectivos, siempre estará presente, por lo que no puede ser despreciable pese a que en proporción sean mayores los aciertos que los errores. La consideración de la diligencia emitida por VioGén como acto administrativo y, consecuentemente, que tanto agresor como víctima puedan interponer recurso, supondría un gran avance en la protección a las víctimas de violencia de género, ya que el objetivo final no es lograr una correcta valoración, sino ofrecer a las víctimas la protección que requieren en atención a su verdadero nivel de riesgo de volver a sufrir una agresión.

Si bien es cierto que existen casos como la SAN 2350/2020, ECLI:ES:AN:2020:2350 o el Dictamen del Consejo de Estado 661/2021, en los que se reconoce la existencia de responsabilidad patrimonial por errores en la valoración del riesgo, debe existir otro procedimiento que no llegue tarde. Es decir, que no se encargue de compensar el daño causado sino de prevenir que se produzca el daño. Cosa que la consideración de la diligencia de VioGén como acto administrativo cumple a la perfección, ya que ofrece la posibilidad de interponer recurso administrativo, el cual es de carácter gratuito ya que no se requiere la intervención de letrado y, posteriormente, si el/la recurrente lo considera oportuno recurso contencioso-administrativo, para poder solventar el error y así aumentar la protección de las víctimas.

Bibliografía

BLASCO DÍAZ, J.L. El sentido de la transparencia administrativa y su concreción legislativa. En: GARCÍA MACHO, R.J., coord. *Derecho administrativo de la información y administración transparente*. Madrid: Civitas, 2010, pp. 143-150.

COMISIÓN EUROPEA. Annex to the Communication to the Commission: Approval of the content of the draft Communication from the Commission - Commission Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689 (AI Act) (C(2025) 924 final). Brussels, 2025.

CORMEN, T., LEISERSON, C., RIVEST, R. y STEIN, C. *Introduction to Algorithms*. Cambridge (MA): MIT Press, 2009.

ESTÉVEZ MENDOZA, L.M. Inteligencia artificial y violencia contra las mujeres: ¿Funcionan los sistemas automatizados de evaluación del riesgo? *PERSPECTIVAS*, Revista de Ciencias Jurídicas y Políticas, 2021, n. 3, pp. 127-141.

GARCÍA DE ENTERRÍA Y MARTÍNEZ-CARANDE, E. y FERNÁNDEZ RODRÍGUEZ, T.R. *Curso de derecho administrativo I*. 20.^a ed. Madrid: Civitas; Thomson Reuters, 2022.

GONZÁLEZ ÁLVAREZ, J.L., LÓPEZ OSSORIO, J.J. y MUÑOZ RIVAS, M. *La valoración policial del riesgo de violencia contra la mujer pareja en España – Sistema VioGén*. Madrid: Secretaría General Técnica, Ministerio del Interior, 2018, pp. 1-120.

LA MONCLOA. Interior recurre a la tecnología de inteligencia artificial para mejorar la valoración policial de riesgo en casos de violencia de género [en línea]. 2020. Disponible en: <https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/interior/Paginas/2020/151220-inteligencia.aspx> [Consultado: 8 septiembre 2025].

LÓPEZ OSSORIO, J.J. (2023, 22 de marzo). Prevención de la violencia de género desde el sistema VioGén. Ponencia en Debates de la UNED, "Violencia de género: lo que nos cuentan y lo que no nos cuentan". Madrid, España.

LÓPEZ OSSORIO, J.J., GONZÁLEZ ÁLVAREZ, J.L. y ANDRÉS PUEYO, A. Eficacia predictiva de la valoración policial del riesgo de la violencia de género. *Psychosocial Intervention*, 2016, vol. 25, n.º 1, pp. 1–7.

LÓPEZ OSSORIO, J.J., MUÑOZ VICENTE, J.M., ANDRÉS PUEYO, A. y PASTOR BRAVO, M. *Guía de aplicación del formulario VFR 5.0-H en la valoración forense del riesgo. Sistema de seguimiento integral en los casos de violencia de género*. Madrid: Ministerio del Interior, 2020, pp. 25–26.

MARTÍNEZ GARAY, L., BOIX PALOP, A., BRIZ REDÓN, A., FLORES GIMÉNEZ, F., GARCÍA ORTIZ, A., MOLINA SÁNCHEZ, M., MONTES SUAY, F., PALMA ORTIGOSA, A., PERIS MANGUILLOT, A. y SORIANO ARANZ, A. *Three predictive policing approaches in Spain: Viogén, RisCanvi and Veripol*. Valencia: Publicaciones Univ. València, 2024.

MOLINA SÁNCHEZ, M. La inteligencia artificial como mecanismo para perennizar la desigualdad. Especial referencia al sistema VioGén. En: RODRÍGUEZ SÁNCHEZ, A., TALAVERA CORDERO, P., DOMÍNGUEZ ÁLVAREZ, J.L. y TERRÓN SANTOS, D., eds. *Desafíos éticos, jurídicos y tecnológicos del avance digital*. Madrid: lustel, 2023, pp. 119–134.

SAN MARTÍN SEGURA, D. Prevención algorítmica de la violencia de género: la discrecionalidad policial como decisión tecnológica en el contexto de VIOGÉN. *Estudios Penales y Criminológicos*, 2023, vol. 44, suplemento, pp. 1-35.

SECRETARÍA DE ESTADO DE SEGURIDAD. *Instrucción 4/2019, por la que se establece un nuevo protocolo para la valoración policial del nivel de riesgo de violencia de género (Ley Orgánica 1/2004), la gestión de la seguridad de las víctimas y seguimiento de los casos a través del Sistema de Seguimiento Integral de los Casos de Violencia de Género (Sistema VioGén)*. Madrid: Ministerio del Interior, 2019a.

SECRETARÍA DE ESTADO DE SEGURIDAD. *Guía de procedimiento VPR 5.0 y VPER 4.1: Protocolo de valoración policial del riesgo y gestión de la seguridad de las víctimas de violencia de género*. Madrid: Ministerio del Interior, 2019b.

SECRETARÍA DE ESTADO DE SEGURIDAD. *Instrucción 5/2021, por la que se establece el protocolo de primer contacto policial con víctimas de violencia de género en situación de desprotección (protocolo cero)*. Madrid: Ministerio del Interior, 2021a.

SECRETARÍA DE ESTADO DE SEGURIDAD. *Instrucción 8/2021, por la que se establecen medidas para prevenir la violencia de género ocasionada por agresores persistentes*. Madrid: Ministerio del Interior, 2021b.

SECRETARÍA DE ESTADO DE SEGURIDAD. *Instrucción 11/2022, por la que se actualizan los procedimientos de actuación en la gestión del riesgo de los denominados “casos resistentes” de violencia de género*. Madrid: Ministerio del Interior, 2022.

SECRETARÍA DE ESTADO DE SEGURIDAD. *Instrucción 1/2023, por la que se desarrolla la obligación de comunicación a la víctima de los antecedentes del agresor en los denominados casos de “agresores persistentes” y otras medidas de protección y seguridad en los casos de violencia de género*. Madrid: Ministerio del Interior, 2023.

SECRETARÍA DE ESTADO DE SEGURIDAD. *Instrucción 1/2025, por la que se establece un nuevo protocolo para la valoración y gestión policial del nivel de riesgo de violencia de género y seguimiento de los casos a través del sistema VioGén-2*. Madrid: Ministerio del Interior, 2025.

SIMÓN CASTELLANO, P. Justicia cautelar e inteligencia artificial: la alternativa a los atávicos heurísticos judiciales. Barcelona: JM Bosch editor, 2021.

SORIANO ARANZ, A. Decisiones automatizadas y discriminación: Aproximación y propuestas generales. *Revista General de Derecho Administrativo*, 2021, n.º 56.

ZURITA BAYONA, J. Violencia contra la mujer: marco histórico evolutivo y predicción del nivel de riesgo. 2014. [Tesis doctoral, Universidad Autónoma de Madrid]. Repositorio institucional Universidad Autónoma de Madrid.

[ACCEDE A LA GRABACIÓN](#)

La Inteligencia Artificial (IA), cuando se aplica con responsabilidad ética y perspectiva humanista, puede ser una poderosa aliada en la promoción de los derechos humanos, especialmente en lo que respecta a la inclusión social y la accesibilidad universal. Esta comunicación examina el impacto positivo de la IA en la vida de personas en situación de vulnerabilidad, con ejemplos reales que demuestran cómo estas tecnologías amplían oportunidades y fortalecen precisamente el ejercicio de los mencionados derechos humanos.

En el ámbito educativo y laboral, la IA permite una formación personalizada que se adapta a las necesidades individuales para que se puedan desarrollar competencias socioemocionales que mejoren la empleabilidad. En la selección de personal, herramientas basadas en IA están ayudando a reducir sesgos, facilitar entrevistas accesibles y valorar las capacidades reales, promoviendo la inclusión laboral de personas con discapacidad.

En términos de accesibilidad, diversas tecnologías ayudan por ejemplo a las personas ciegas a identificar objetos, leer textos en tiempo real o incluso conocer el contenido de una carta en un restaurante mediante el escaneo con el móvil. Además, asistentes virtuales como Alexa incrementan la autonomía en la vida cotidiana, permitiendo, mediante comandos de voz, gestionar tareas, acceder a información o controlar dispositivos domésticos.

¹³⁹ El presente trabajo se enmarca en el proyecto de investigación “la planificación administrativa, ref. PID2023-152781NB-I00” que se encuadra en el Plan Nacional I+D+i: PLANADM 2024-2026, en el proyecto “Algorithmic Law”, ref. CIPROM/2024/16, que se enmarca en el programa Prometeu Generalitat Valenciana: AILAW, 2025-2026 y del convenio C036/23-OT entre la entidad pública empresarial Red.es agrupación de entidades formada por la Universidad del País Vasco/EUSKAL HERRIKO UNIBERSITATEA (UPH/EHU), la Universitat de València, la Fundación Instituto de Investigación Sanitaria “Marqués de Valdecilla”, la Entidad de Gestión de Derechos de los Productores Audiovisuales LA (EGEDA), la Universitat Rovira i Virgili, y la Fundación Hay Derecho, para impulsar la implementación de la carta de derechos digitales en el ámbito de entornos específicos.

Paralelamente, la IA está siendo utilizada tanto para mejorar la accesibilidad digital de sitios web como para llevar a cabo la adaptación arquitectónica de espacios físicos, eliminando barreras y favoreciendo la participación plena.

En definitiva, esta ponencia defiende la IA como un motor de transformación social que, aplicada con principios de justicia y accesibilidad, contribuye a construir entornos más inclusivos, equitativos y respetuosos con la diversidad.

En los últimos años, la Inteligencia Artificial (IA) se ha consolidado como un verdadero motor de cambio social. Más allá de los discursos distópicos que dominan con frecuencia el debate público, la realidad muestra un horizonte distinto: la IA está contribuyendo a la ampliación de derechos, al fortalecimiento de la dignidad y a la inclusión de colectivos históricamente marginados. Su capacidad para analizar grandes volúmenes de datos, anticipar riesgos y activar respuestas preventivas ofrece un potencial inédito para reducir la desigualdad y garantizar el acceso a derechos básicos.

En España, por ejemplo, algunos ayuntamientos han incorporado sistemas predictivos que identifican con antelación el riesgo de desahucio cruzando datos de impagos, desempleo y variaciones familiares. Esta anticipación permite a los servicios sociales intervenir antes de que se produzca la pérdida de la vivienda. Del mismo modo, programas de reparto de ayuda alimentaria gestionados con algoritmos de IA ajustan en tiempo real la distribución de recursos, evitando demoras y burocracias que dejan fuera a las personas más vulnerables.

Algunos de esos proyectos de implantación de Inteligencia Artificial en la gestión social se han identificado especialmente en Canarias, donde hasta 50 ayuntamientos han participado en experiencias de Laboratorios de Innovación con IA, según la Fundación Emprende (<https://www.fundacionemprende.org/familias-riesgo-inteligencia-artificial-ayuntamientos/>).

En el ámbito internacional, el caso de Estonia resulta paradigmático: su gobierno está utilizando la IA generativa para diseñar programas de bienestar social personalizados. No se trata ya de políticas generalistas, sino de intervenciones adaptadas a cada ciudadano que optimizan la distribución de recursos y reducen la desigualdad. En paralelo, organizaciones humanitarias como el Programa Mundial de Alimentos emplean IA para planificar rutas de distribución de víveres en crisis de refugiados o en territorios afectados por catástrofes, reduciendo desperdicios y garantizando rapidez en la ayuda (https://www.acta.es/medios/articulos/cultura_y_sociedad/226001.pdf).

Por otro lado, la Inteligencia Artificial está desempeñando un papel decisivo en la reducción de desigualdades en el acceso a la salud que es un aspecto central cuando hablamos de garantizar este derecho fundamental. Gracias a la IA, es posible acercar diagnósticos y tratamientos a personas que viven en zonas rurales o que carecen de recursos para desplazarse a grandes centros hospitalarios.

En España, la telemedicina apoyada por algoritmos inteligentes ya se utiliza en comunidades autónomas con alta dispersión geográfica, como Castilla y León o Galicia, donde los sistemas de apoyo al diagnóstico permiten que médicos de atención primaria en pequeños municipios puedan contar con una segunda opinión basada en datos de miles de pacientes, sin necesidad de derivar a los enfermos a hospitales de referencia situados a cientos de kilómetros.

En el ámbito europeo, se han puesto en marcha proyectos piloto de cribado remoto de enfermedades cardiovasculares y respiratorias, en los que dispositivos conectados recopilan información del paciente y sistemas de IA analizan esos datos para detectar posibles riesgos. Esto resulta especialmente relevante en regiones con población envejecida o con limitaciones de transporte sanitario, porque garantiza que personas que de otro modo tendrían más dificultades reciban una atención preventiva y temprana.

Además, la IA aplicada a la gestión sanitaria permite optimizar la asignación de recursos, de modo que hospitales públicos puedan organizar mejor las listas de espera y priorizar los casos más urgentes. Esta dimensión organizativa también tiene un impacto directo en la equidad, ya que evita que los recursos se concentren solo en las grandes ciudades y facilita que pacientes de entornos menos favorecidos accedan a servicios de calidad en tiempos razonables.

En países del Sur Global la Inteligencia Artificial se está convirtiendo en una aliada imprescindible por lo que respecta al derecho a la salud. Un ejemplo significativo es el uso de aplicaciones móviles con sistemas de diagnóstico asistido por IA en regiones de África subsahariana para la detección precoz de la malaria o de enfermedades respiratorias infantiles. Estas herramientas permiten que personal sanitario con una formación básica pueda obtener diagnósticos rápidos y fiables, reduciendo la dependencia de especialistas que a menudo se concentran en grandes ciudades y están fuera del alcance de comunidades alejadas. En esta línea, Uganda y Nigeria han impulsado proyectos que combinan pruebas rápidas con algoritmos de aprendizaje automático, con resultados que han mejorado tanto la rapidez en la intervención como la tasa de supervivencia.

El campo de la salud mental ha sido otro beneficiario del avance de la IA. Sistemas en forma de chatbots terapéuticos y aplicaciones móviles proporcionan apoyo instantáneo a personas en crisis emocional, ofreciendo escucha activa y pautas de afrontamiento. Esta accesibilidad, especialmente relevante en lugares donde la atención psicológica presencial es insuficiente, ha reducido estigmas, mejorado la adherencia a tratamientos y facilitado el diagnóstico precoz de trastornos mentales. El resultado es un refuerzo del derecho a la salud y a la protección integral, especialmente en zonas rurales o contextos de exclusión.

LA IA y las personas mayores

La Inteligencia Artificial está transformando también el apoyo a las personas mayores, colectivo que a menudo sufre aislamiento, soledad no deseada y riesgo de caídas. A modo de proyectos piloto, en municipios españoles han instalado sensores inteligentes en domicilios de mayores que viven solos. Estos dispositivos, a través de IA, monitorizan actividad diaria, detectan cambios en rutinas y envían alertas automáticas a familiares o cuidadores si el sistema percibe anomalías que puedan indicar una emergencia. De este modo, no solo se previenen accidentes

graves, sino que se refuerza la autonomía personal, facilitando que las personas mayores prolonguen su vida independiente y fortalezcan su derecho a la dignidad y a la salud.

Por lo que respecta al uso de los asistentes virtuales basados en IA, el programa de la Cruz Roja con Alexa es una innovadora iniciativa destinada a mejorar la autonomía, la comunicación y la calidad de vida de las personas mayores y otros colectivos vulnerables a través del uso de asistentes de voz inteligentes. Este proyecto nació durante el confinamiento de la pandemia de la Covid-19, cuando la Cruz Roja Española, junto a organizaciones como la Fundación Tecnologías Sociales, Accenture y Amazon España, desarrolló en tiempo récord una “skill” (utilidad o aplicación) específica para los dispositivos Alexa. El objetivo principal era aportar compañía, acceso a información útil -como consejos de prevención del coronavirus y primeros auxilios: También se perseguía facilitar servicios de acompañamiento virtual, especialmente para quienes experimentaban soledad o aislamiento social debido a la brecha digital (<https://www2.cruzroja.es/web/ahora/-/dispositivos-de-voz-para-favorecer-la-autonom-c3-ada-y-reducir-la-soledad-de-26.000-personas-mayores>).

La skill de Cruz Roja para Alexa sirve para que los usuarios puedan a través de ella y de forma sencilla contactar con la organización. También pueden recibir información relevante, activar recordatorios de medicación o citas, hacer donaciones, acceder a vídeos guiados de ejercicios físicos adaptados, reducir el estrés y formarse sobre salud y autocuidado. Además, facilita la conexión con familiares y voluntarios, y realiza servicios como alarmas y la organización de la vida cotidiana. Se han implementado funciones visuales y multimedia, como ilustraciones para formación en primeros auxilios y tutoriales en vídeo diseñados para personas mayores.

El impacto de esta tecnología de voz ha resultado muy positivo: más del 69% de los participantes manifestó sentirse más acompañados, mejorar su comunicación social y su entretenimiento, y experimentar mayor organización diaria y autonomía personal, además de aliviar la soledad no deseada. El programa, -que ya atiende a más de 26.000 personas mayores-, ha sido especialmente útil para quienes tienen discapacidad o limitaciones de movilidad, pues permite pedir ayuda y gestionar contactos de emergencia simplemente con comandos de voz. Como parte del desarrollo y análisis de la iniciativa, se han distribuido dispositivos Alexa en hogares y residencias por toda España, beneficiando ya a decenas de miles de usuarios en riesgo de aislamiento social.

Paralelamente, y también en cuanto a lo que avanza la lucha contra la soledad con la ayuda de la Inteligencia Artificial, han surgido otras herramientas como Onorato AI (<https://onoratoai.com>). Se trata de un loro virtual dotado de IA que está siendo utilizado como compañía para mayores en riesgo de aislamiento. El sistema simula conversaciones, detecta estados de ánimo mediante el lenguaje y motiva a mantener vínculos sociales. Estas tecnologías

contribuyen al derecho al bienestar, previniendo efectos negativos de la soledad en la salud mental y física.

A lo anterior hay que añadir las experiencias internacionales que incluyen robots sociales en residencias (por ejemplo, en Japón o Corea del sur) que reconocen rostros, voces, permiten videollamadas, dan noticias, reproducen música, etc., reduciendo la sensación de soledad. En particular, proyectos como “Paro”, el robot foca japonés, han demostrado beneficios en salud mental, aunque no equivale a una persona, su uso terapéutico con IA controla estímulos sensoriales que sirven para generar calma.

Cómo la IA está mejorando la accesibilidad universal

La Inteligencia Artificial está transformando profundamente la manera en que la sociedad aborda la accesibilidad tanto en entornos digitales como físicos, convirtiéndose en una herramienta clave para la eliminación de barreras y la promoción de la participación plena de todas las personas. Su integración en distintos ámbitos ha dejado de ser una promesa de futuro para materializarse en resultados concretos que benefician a millones de personas con discapacidad, personas mayores y a colectivos con dificultades de acceso o comprensión, permitiendo la igualdad y el ejercicio efectivo de derechos.

La accesibilidad ha dado un salto cualitativo gracias a la IA aplicada en el ámbito de la discapacidad. Personas ciegas o con movilidad reducida son capaces de moverse con mayor seguridad por entornos urbanos gracias al desarrollo de aplicaciones móviles de navegación inteligente y sistemas de reconocimiento de entorno que describen en tiempo real lo que sucede alrededor del usuario. Los asistentes de voz y traductores automáticos de lenguaje braille o lengua de signos facilitan la inclusión educativa y laboral, haciéndolo posible que personas con discapacidad accedan a contenidos y procesos de selección de empleo sin discriminación. Entidades como la ONCE están impulsando la adaptación de estas tecnologías a contextos locales, asegurando que respondan a barreras específicas y expandiendo el derecho a la igualdad de oportunidades.

Junto a las diferentes iniciativas públicas hay empresas privadas como Samsung que también están poniendo en marcha iniciativas que aúnan el uso de tecnología punta basada en IA con la atención a las personas más vulnerables. En España, sin ir más lejos, esta empresa ha desarrollado la herramienta “Unfear” que permite modular sonidos ambientales para personas con autismo y sensibilidad auditiva, previniendo crisis y garantizando la plena participación en la vida pública (<https://www.samsung.com/es/tecnologiaconproposito/accesibilidad-bienestar/unfear/>).

Además, en el entorno digital, la IA actúa como catalizadora de la accesibilidad web. Gracias a ella las diferentes páginas web y las aplicaciones digitales son usables por personas con

dificultad visual, auditiva, cognitiva o motriz. Ejemplos concretos de soluciones basadas en IA, como [accessiBe](https://accessibe.com) (<https://accessibe.com>) y [UserWay](https://userway.org) (<https://userway.org>), demuestran la capacidad de los algoritmos para analizar el código y los elementos de una web en tiempo real, adecuando automáticamente contrastes de color, navegación por teclado y compatibilidad con lectores de pantalla. Estas plataformas detectan automáticamente elementos problemáticos y los ajustan sin intervención humana directa, lo que reduce el margen de error y mantiene la conformidad con estándares internacionales como las WCAG y la ADA norteamericana.

La personalización es otro avance reseñable: widgets o pequeñas aplicaciones inteligentes permiten que cada usuario adapte la experiencia digital según necesidades concretas, como perfiles especializados para personas daltónicas, con baja visión o dislexia. En grandes organizaciones, el impacto de soluciones como UserWay ha sido medido: tasas de rebote más bajas y mayor satisfacción, pero, sobre todo, la certeza de que millones de usuarios pueden acceder a información y servicios con autonomía y dignidad. La Universidad de Alicante, por ejemplo, ha implementado proyectos de accesibilidad digital basados en IA para fomentar la autonomía y participación universitaria de personas con cualquier grado de discapacidad.

En el ámbito del contenido, tecnologías de procesamiento de lenguaje natural generan descripciones automáticas de imágenes, aseguran que todo contenido multimedia tenga subtítulo en tiempo real y convierten documentos complejos en versiones de lectura más fácil. Las herramientas “Seeing AI” de Microsoft, “Google Lookout” o [Read Easy.ai](https://readeasy.ai) están permitiendo a personas con discapacidad visual “leer” su entorno, identificar billetes o etiquetas en comercios y comprender menús o señalética en espacios públicos, disminuyendo drásticamente la dependencia de terceros para actividades cotidianas.

En la comunicación audiovisual, la Inteligencia Artificial ofrece soluciones a personas con discapacidad auditiva mediante transcripción automática y generación de subtítulos en vídeos en tiempo real, facilitando así su inclusión en entornos educativos, laborales y en el acceso a contenidos culturales o informativos. Aplicaciones como Google Live Transcribe o subtítulos automáticos en plataformas como YouTube son ya imprescindibles para millones de personas, y han demostrado un impacto positivo en la integración y reducción de la brecha de acceso al conocimiento.

La eliminación de barreras propiciada por la IA va más allá del puro cumplimiento legal, promoviendo valores de autonomía, igualdad y dignidad. Las tecnologías inteligentes permiten adaptar no solo el entorno construido sino también la información, la comunicación y el acceso a la participación en todos los ámbitos sociales. La generación automática de descripciones de imágenes en redes sociales, traducidas en tiempo real, hace que el activismo, la cultura y la información estén realmente al alcance de todos, y no reservados a quienes no enfrentan dificultades sensoriales o tecnológicas.

El impacto de la IA en el ámbito digital va más allá de lo técnico: organizaciones como Iluion Accesibilidad ofrecen auditorías y soluciones específicas para la adaptación de sitios web a las normativas, así como asesoría para crear estrategias de comunicación social inclusiva. El reto es que la información llegue realmente a todos los públicos, en todos sus formatos y soportes, sin barreras derivadas de la incapacidad técnica o el desconocimiento normativo.

El desarrollo de la accesibilidad física tampoco ha sido ajeno a la revolución impulsada por la IA. La adaptación arquitectónica y la eliminación de barreras espaciales han dado un salto cualitativo gracias al uso de algoritmos avanzados en el diseño, manejo, control y mantenimiento inteligente de edificaciones. La visión por ordenador permite crear sistemas en los que sensores inteligentes reconocen la presencia y necesidades de usuarios con movilidad reducida o dispositivos de asistencia, activando por ejemplo ascensores sin necesidad de pulsar botones o ajustando la velocidad y tiempo de cierre de puertas automáticas.

Un caso concreto es el del PowerhouseBrattørkaia en Noruega, donde la Inteligencia Artificial gestiona la iluminación ajustando la entrada de luz natural respecto a la ocupación real del espacio, y adaptando parámetros para que todo el edificio sea eficiente a cualquier usuario, incluido quienes necesiten condiciones de luminosidad específicas para desplazarse o trabajar. Este tipo de soluciones, basadas en el llamado “diseño paramétrico”, permiten a arquitectos crear formas y estructuras adaptativas, atendiendo tanto a la complejidad geométrica como a la respuesta personalizada a las necesidades de cada colectivo.

En el caso de la movilidad vertical, la Inteligencia Artificial ha hecho posible que los ascensores actuales cuenten con algoritmos de control inteligente, capaces de detectar la presencia de usuarios con silla de ruedas, priorizar paradas o anticipar la demanda en picos de uso, evitando esperas y facilitando la autonomía en edificios públicos y privados. Interfaces de usuario apoyadas en IA, a través del reconocimiento de voz o gestos, han facilitado el uso de ascensores a personas con discapacidad visual o motora, eliminando obstáculos y promoviendo independencia real en el entorno urbano y doméstico.

El mantenimiento predictivo, otra aportación de la IA, posibilita que los sistemas de transporte vertical estén siempre en condiciones óptimas de uso, minimizando averías y tiempos de inactividad. Para las personas que dependen del ascensor para desplazarse, este avance supone una mejora inestimable en su calidad de vida y autonomía diaria; cualquier interrupción en el servicio puede suponer una exclusión súbita y dramática, por lo que la fiabilidad técnica es ahora una auténtica garantía de derecho a la movilidad.

La aplicación de algoritmos inteligentes permite analizar flujos de tráfico peatonal en grandes espacios públicos, identificando puntos de congestión y barreras arquitectónicas “invisibles”, que a menudo pasan desapercibidas en auditorías tradicionales. Así, se pueden diseñar rutas accesibles, señaléticas adaptadas y mecanismos de alerta automática ante obstáculos,

promoviendo un modelo de “diseño universal” en lugar de la clásica adaptación exclusiva para personas con discapacidad. Plataformas como RoomGPT (<https://www.roomsgpt.io/es>), asistidas por IA, ayudan tanto a arquitectos como a usuarios finales a planificar reformas domiciliarias y de oficinas, simulando escenarios y proponiendo soluciones adaptadas a necesidades concretas, con el consiguiente ahorro de tiempo y costes y mayor participación de quienes realmente utilizarán esos espacios.

Ejemplos recientes incluyen la colaboración de aseguradoras y promotoras inmobiliarias para usar Inteligencia Artificial en el ajuste dinámico de alturas de mostradores, anchura de pasillos y control de obstáculos, a partir de datos recogidos en tiempo real sobre usuarios con muletas, andadores o portadores de prótesis robóticas. Estas mejoras, además de incrementar la seguridad y comodidad, redundan en beneficios para todo el público: espacios sin barreras son más cómodos para familias con carritos, personas mayores o temporales convalecientes, fortaleciendo la idea de que la accesibilidad universal beneficia realmente al conjunto de la sociedad.

Desde la perspectiva laboral, la IA ha impulsado avances en la accesibilidad de los espacios de trabajo, mediante programas de formación personalizados, asistentes virtuales que ayudan a la navegación dentro de oficinas y fábricas, transcripción automática de reuniones y adaptación del mobiliario en función del análisis de posturas y trayectorias de usuarios. Herramientas como InclunIA de Fundación ONCE analizan datos de candidatos y ofertas para promover la igualdad de oportunidades y adecuación de los espacios a las personas con discapacidad, derribando barreras invisibles que frenaban el acceso y promoción profesional.

No menos relevante es el vínculo que existe entre accesibilidad digital y accesibilidad física fomentado por la IA. El diseño universal asistido por algoritmos permite integrar los requisitos legales locales e internacionales (por ejemplo, WCAG para lo digital y normativa de supresión de barreras arquitectónicas para lo físico) en un mismo flujo de trabajo. Plataformas que analizan automáticamente los planos de edificios y comparan con la normativa vigente alertan a los técnicos de posibles incumplimientos, sugieren soluciones y desarrollan iteraciones en tiempo récord, acelerando el proceso de obtención de licencias de accesibilidad y mejorando exponencialmente la calidad final de los proyectos.

En otro orden de cosas, la selección de personal con apoyo de IA ha aterrizado en el ámbito de la empresa social, permitiendo valorar competencias profesionales de los candidatos de forma objetiva y sin sesgos humanos. Este avance ha facilitado la inclusión de colectivos que históricamente sufrían discriminación en el acceso al empleo, como personas con discapacidad, migrantes o mayores de 45 años, contribuyendo al derecho al trabajo y a la igualdad de trato.

Derecho a la educación

La Inteligencia Artificial se está convirtiendo en una herramienta esencial para estudiantes con dificultades de aprendizaje. Existen ya sistemas que identifican la dislexia o el déficit de atención en etapas tempranas, también hay plataformas que adaptan ejercicios y alertan a los docentes para aplicar adaptaciones específicas. La IA favorece la inclusión y posibilita un acceso real y una permanencia efectiva en el sistema educativo, superando barreras históricas que excluyeron a muchos alumnos.

Por ejemplo, Khanmigo, de Khan Academy en colaboración con OpenAI, es un asistente virtual basado en GPT-4 que guía al estudiante en la resolución de problemas mediante preguntas y sugerencias. Permite diseñar itinerarios de aprendizaje, ajustar el contenido al nivel del alumno, preparar y corregir ejercicios, y también apoyar a los profesores en la elaboración de material y evaluación. (<http://openai.com/index/khan-academy>). Además, Khanmigo ofrece diversas utilidades gratuitas para docentes en Estados Unidos, gracias a una colaboración con Microsoft que da acceso a su plataforma sin coste para maestros de educación básica. (<http://cnbc.com/2024/05/21/microsoft-khan-academy-launch-free-ai-assistant-for-all-us-teachers.html>).

Otra herramienta relevante es Read Easy.ai, que facilita la lectura de textos complejos simplificándolos, adaptando su vocabulario y ajustando su nivel de legibilidad. Cuenta con extensiones para navegadores, complementos para procesadores de texto y API para desarrolladores. (<http://readeasy.ai/support>). Por su parte, DyslexiaBuddy aporta un acompañamiento personalizado: permite leer textos mediante voz, ajustar fuentes y colores, y ofrece un tutor virtual para resolver dudas al instante. (<http://dyslexiabuddy.com>). Otra propuesta es Lexy, que mediante actividades multisensoriales -como reconocimiento de sonidos, escritura táctil y juegos interactivos- mejora la fluidez lectora, ortográfica y la comprensión en quienes tienen dislexia. (<http://dyslexia.ai>).

Estas tecnologías no solo adaptan contenidos al nivel de comprensión lectora o cognitiva del estudiante, sino que detectan barreras lingüísticas o visuales en tiempo real. Al transformar los materiales didácticos con fuentes legibles, fondos adecuados, espacios entre letras o líneas suficientes, y vocabulario accesible, se garantiza que ningún estudiante quede descartado por la forma -y no por la calidad- del contenido educativo.

De este modo, la IA contribuye directamente al derecho a la educación y a la igualdad de oportunidades. Contribuye a que el alumnado con necesidades diversas acceda a la misma calidad de aprendizaje que quienes no enfrentan esas barreras. Al mismo tiempo supone un apoyo significativo para docentes, pues optimiza los plazos de ejecución del diseño, corrección y seguimiento, permitiéndoles concentrarse en aspectos pedagógicos más finos, como motivar

a quienes están menos involucrados o facilitar el progreso acelerado de quienes tienen mayor facilidad.

A la hora de destacar ejemplos en España, hay que citar una aplicación denominada Dyetective, creada por Change Dyslexia. Emplea aprendizaje automático y juegos digitales para detectar riesgos de dislexia. Ofrece además más de 40.000 ejercicios lúdicos personalizados para quienes ya han sido diagnosticados, para trabajar lectura o escritura. Ha sido probada en comunidades como Madrid o Murcia. (<http://dataviz.unesco.org/es/articles/una-aplicacion-espanola-basada-en-la-ia-para-ayudar-detectar-la-dislexia-gana-un-premio-de-la-unesco>).

En la misma línea Neurekalab, una startup catalana, ha creado una herramienta de IA para detectar de forma temprana dislexia, discalculia y TDAH en edades de 5 a 8 años. Se usa en aulas durante lecciones de lengua y matemáticas, utilizando tabletas u ordenadores, con tareas breves diarias, lo que permite generar al final de curso un perfil cognitivo detallado.

La Universidad Oberta de Catalunya (UOC), por su parte, está desarrollando un programa interactivo personalizado que utiliza tecnologías, movimiento corporal e IA para trabajar la conciencia fonológica en menores de tres a siete años. El objetivo es ayudar a maestros, logopedas y otros profesionales a intervenir temprano ante dificultades iniciales de lectura y lenguaje. (<http://uoc.edu/es/news/2023/002-lectura-menores-IA>).

El proyecto Aula IA de la Universidad de Alicante pone a disposición de más de 400 escolares y decenas de profesores materiales didácticos en educación primaria que introducen conceptos de la IA (su funcionamiento, implicaciones éticas, impacto social). Se han hecho talleres en varios municipios para alfabetización tecnológico-científica. (<http://alicantepress.com/art/49049/la-universidad-de-alicante-desarrolla-un-proyecto-para-introducir-la-ia-en-los-colegios>).

En el CEPA Daoíz y Velarde (Madrid), un centro de educación de adultos se ha introducido la IA en la gestión institucional, en la elaboración de planes de absentismo, en la actualización de protocolos y en herramientas que apoyan el aprendizaje, tanto para docentes como estudiantes adultos. (<http://epale.ec.europa.eu/es/content/innovacion-educativa-con-ia-en-el-cepa-daoiz-y-velarde>).

Conclusión

No obstante, los desafíos que plantea la expansión de la Inteligencia Artificial no son menores. Resulta imprescindible garantizar la protección de los datos personales, prevenir la reproducción de sesgos algorítmicos y asegurar que los sistemas reflejen la diversidad real de la sociedad. Solo de esta manera se evitará que la accesibilidad ofrecida por la IA se convierta en un patrón homogéneo que perpetúe las mismas exclusiones históricas que se pretende superar. La interoperabilidad entre plataformas, el acceso económico a las soluciones

tecnológicas y la capacitación técnica de quienes diseñan e implementan estos sistemas son también cuestiones urgentes que requieren una respuesta colectiva y sostenida en el tiempo.

En conclusión, la Inteligencia Artificial no debe entenderse únicamente como un conjunto de herramientas de gran valor técnico, sino como un auténtico motor de transformación social y de promoción de los derechos humanos. La IA está contribuyendo a configurar un modelo educativo y social basado en el diseño universal, la accesibilidad y la participación plena. Los ejemplos ya existentes —desde tutores virtuales que eliminan barreras de aprendizaje, hasta asistentes de navegación que facilitan la autonomía en la web, pasando por sistemas de movilidad inteligente y recursos didácticos adaptados en tiempo real— demuestran que el futuro de la accesibilidad no es una aspiración lejana, sino una realidad en construcción. Con un uso responsable y ético, la IA se convierte en aliada de la inclusión social y en garante del derecho a participar en igualdad de condiciones en todos los ámbitos de la vida.

Bibliografía

AMI. Impacto de la Inteligencia Artificial en los derechos de las personas con discapacidad [en línea]. Disponible en: <https://www.consaludmental.org/publicaciones/Impacto-inteligencia-artificial-derechos-discapacidad.pdf>

Ascensores en Sevilla. Cómo la IA puede ayudar a eliminar barreras arquitectónicas [en línea]. Disponible en: <https://ascensoresensevilla.com/blog/como-la-inteligencia-artificial-puede-ayudar-a-eliminar-barreras-arquitectonicas/>

BibLus. IA en arquitectura: 8 ejemplos de aplicación [en línea]. 11 mayo 2025. Disponible en: <https://biblus.accasoftware.com/es/ai-en-la-arquitectura-8-ejemplos-de-aplicacion/>

Bizkaia. Inteligencia Artificial para las personas mayores [en línea]. Disponible en: <https://www.bizkaia.eus/documents/9027320/11569571/%2307+Inteligencia+Artificial+VF-ES.pdf>

Cruz Roja Española. Inteligencia Artificial como Tecnología Humanitaria para un mundo mejor [en línea]. Disponible en: <https://www2.cruzroja.es/web/ahora/-/inteligencia-artificial-como-tecnologia-humanitaria-para-un-mundo-mejor>

Documentación Social. Inteligencia Artificial: una herramienta clave en la lucha contra la pobreza y la exclusión social [en línea]. 8 julio 2024. Disponible en: <https://documentacionsocial.es/17/accion-social/inteligencia-artificial-una-herramienta-clave-en-la-lucha-contra-la-pobreza-y-la-exclusion-social/>

FECYT. La Inteligencia Artificial y su impacto en el bienestar social [en línea]. Disponible en: https://www.fecyt.es/sites/default/files/2025-02/informe_tendencias_2024_la_inteligencia_artificial_y_su_impacto_en_el_bienestar_social_0.pdf

Fundación ONCE. IA y soledad no deseada [en línea]. Disponible en: https://accessibilas.es/wp-content/uploads/2025/03/20250107_Fundacion-Once_Informe-Estudio-IA-y-la-soledad-no-deseada_final.pdf

Fundación ONCE. La IA, una oportunidad para el futuro de la inclusión laboral [en línea]. 17 marzo 2025. Disponible en: <https://iamasigual.eu/la-ia-supone-una-oportunidad-para-transformar-el-futuro-de-la-inclusion-laboral/>

GEZE. Eliminación de barreras arquitectónicas en edificios [en línea]. Disponible en: <https://www.geze.es/es/descubrir/temas/eliminacion-de-barreras-arquitectonicas>

Ilunion. Inteligencia Artificial y Accesibilidad: Revolución para la Inclusión [en línea]. 6 febrero 2025. Disponible en: <https://www.ilunion.com/es/blog-puntoilunion/inteligencia-artificial-accesibilidad>

Infocop. Beneficios y riesgos de la IA en la atención a la salud mental [en línea]. 29 abril 2025. Disponible en: <https://www.infocop.es/beneficios-y-riesgos-de-la-ia-en-la-atencion-a-la-salud-mental-segun-mhe/>

La Rioja. Salud utiliza un innovador proyecto basado en IA para mejorar las coberturas de vacunación [en línea]. 12 agosto 2025. Disponible en: <https://actualidad.larioja.org/noticia?n=not-salud-utiliza-un-innovador-proyecto-basado-en-ia-para-mejorar-las-coberturas-de-vacunacion-fre>

Revista Economía Industrial. Impacto de la Inteligencia Artificial en el bienestar social. Ministerio de Industria [en línea]. Disponible en: <https://www.mintur.gob.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/423/SA%CC%81NCHEZ%20ROSADO%20Y%20DI%CC%81EZ%20PARRA.pdf>

RoomGPT. RoomGPT: Diseño de Espacios con Inteligencia Artificial [en línea]. Disponible en: <https://cronuts.digital/es/roomgpt-diseno-espacios-inteligencia-artificial/>

Servimedia. Samsung destaca el poder de la Inteligencia Artificial para ayudar a colectivos vulnerables [en línea]. 30 noviembre 2024. Disponible en: <https://www.servimedia.es/noticias/samsung-destaca-poder-inteligencia-artificial-para-ayudar-colectivos-vulnerables/1411401173>

Solidaridad Intergeneracional. Onorato AI, el loro con Inteligencia Artificial que ayuda a los mayores [en línea]. 26 febrero 2025. Disponible en: <https://solidaridadintergeneracional.es/wp/onorato-ai-el-loro-con-inteligencia-artificial-que-ayuda-a-los-mayores/>

UiX. Diseño inclusivo en la era de IA: accesibilidad digital sin límites [en línea]. 24 abril 2025. Disponible en: <https://uixdesign.com/blog/disenio-inclusivo-ai>

Unite.AI. Las 10 mejores herramientas de accesibilidad con IA para sitios web [en línea]. 31 agosto 2025. Disponible en: <https://www.unite.ai/es/ai-accessibility-tools/>

Universidad de Alicante. La UA se alía con la Inteligencia Artificial para avanzar en accesibilidad web e inclusión [en línea]. 14 mayo 2025. Disponible en: <https://web.ua.es/es/actualidad-universitaria/2025/mayo2025/12-18/la-ua-se-alia-con-la-inteligencia-artificial-para-avanzar-en-accesibilidad-web-e-inclusion.html>

Carla de Paredes Gallardo, Laura Lopez-Almansa Beaus y Susana Berrocal Diaz.
Delitos de odio y discurso violento en Redes Sociales

Carla de Paredes Gallardo
Laura Lopez-Almansa Beaus
Susana Berrocal Diaz
Universidad Europea de Valencia
Facultad de Ciencias Sociales

[ACCEDE A LA GRABACIÓN](#)

Resumen:

En las últimas décadas, las redes sociales han revolucionado la comunicación, permitiendo que cualquier persona difunda mensajes a gran escala sin mediación editorial. Esta democratización de la libertad de expresión, reconocida constitucionalmente, ha generado oportunidades para la participación ciudadana, pero también ha facilitado la proliferación de discursos violentos, discriminatorios y de odio, especialmente dirigidos contra grupos vulnerables. Estas manifestaciones pueden incitar al hostigamiento o amenazas, y en ciertos casos, constituir delitos de odio tipificados penalmente.

Desde una perspectiva teórica y normativa, el discurso de odio se define como cualquier forma de comunicación que ataca o discrimina a personas o grupos por motivos como religión, etnia, género u orientación sexual. No se trata solo de expresiones polémicas, sino de contenidos que alimentan climas de hostilidad y violencia estructural. A nivel global, se ha observado que el discurso de odio en línea puede preceder episodios de violencia real, como genocidios o ataques contra minorías.

Las redes sociales, por su diseño, favorecen la rápida difusión de estos discursos. Factores como el anonimato, los algoritmos de recomendación, la cultura de la inmediatez y el efecto de cámara de eco contribuyen a la normalización de prejuicios y a la desensibilización frente a contenidos discriminatorios. Las víctimas sufren consecuencias psicológicas y sociales, y el daño puede extenderse a comunidades enteras.

España ha respondido con reformas legales, como la del artículo 510 del Código Penal, que sanciona conductas que inciten al odio. Sin embargo, la aplicación práctica enfrenta desafíos: delimitar la incitación, probar la motivación discriminatoria y atribuir responsabilidad en

entornos digitales. Jurisprudencia reciente ha intentado clarificar estos límites, pero persisten debates doctrinales sobre el equilibrio entre libertad de expresión y protección frente al odio. El presente trabajo analiza este fenómeno desde una perspectiva interdisciplinar, proponiendo soluciones éticas y tecnológicas que garanticen la protección efectiva de los derechos fundamentales

Palabras clave: inteligencia artificial, discurso de odio, redes sociales, derechos humanos, libertad de expresión.

1. Introducción

La proliferación del discurso de odio y la violencia digital en redes sociales ha generado una creciente preocupación en el ámbito jurídico y tecnológico. En este contexto, la inteligencia artificial (IA) se presenta como una herramienta clave para la detección, análisis y regulación de estos fenómenos. España ha sido uno de los países que ha impulsado iniciativas relevantes en este campo, buscando integrar soluciones automatizadas en el sistema penal.

La detección y regulación penal del discurso de odio y la violencia digital en redes sociales mediante inteligencia artificial (IA) constituye un ámbito en constante evolución, especialmente en España, donde se han impulsado iniciativas relevantes. Las herramientas basadas en IA permiten monitorizar redes sociales en tiempo real, identificar perfiles que promueven el odio y detectar víctimas, además de generar contranarrativas automáticas eficaces (Gorwa, Binns & Katzenbach, 2020).

Uno de los principales desafíos técnicos radica en la capacidad de los algoritmos para diferenciar entre lenguaje ofensivo y discurso de odio. Esta distinción exige un análisis contextual que considere la intención del emisor, el entorno sociocultural y la identidad del receptor (MacAvaney et al., 2019). La precisión en esta diferenciación es esencial para evitar censuras injustificadas que vulneren el derecho a la libertad de expresión.

Desde una perspectiva legal y ética, el uso de IA plantea interrogantes sobre la validez jurídica de las pruebas generadas automáticamente, la transparencia de los algoritmos y el derecho a la defensa frente a acusaciones basadas en sistemas automatizados. La delgada línea entre discurso de odio y libertad de expresión exige una regulación cuidadosa que garantice el respeto a los derechos fundamentales.

En España, la Oficina Nacional de Lucha contra los Delitos de Odio (ONDOD) trabaja en la integración de estas tecnologías en el sistema penal. No obstante, aún se requiere un marco normativo claro que defina los límites, garantías procesales y condiciones de uso judicial de la IA. La implementación efectiva de estas herramientas debe equilibrar la protección frente al odio digital con la preservación de los derechos democráticos, en especial la libertad de expresión.

2. Fundamentos teóricos y marco teórico del discurso de odio

El discurso de odio puede definirse como toda expresión que promueve, incita o justifica el odio, la discriminación o la violencia contra individuos o grupos por motivos de raza, religión, orientación sexual, género, nacionalidad, discapacidad, ideología u otras características identitarias. Esta definición, aunque ampliamente aceptada, varía según el contexto jurídico, cultural y político, lo que complica su delimitación precisa y su aplicación práctica.

Desde una perspectiva sociológica, el discurso de odio no solo refleja prejuicios individuales, sino que reproduce estructuras de poder y exclusión social. Pierre Bourdieu lo vincula con la violencia simbólica, entendida como aquella que se ejerce a través del lenguaje y las representaciones, actuando como un mecanismo de dominación. De este modo, el discurso de odio refuerza estigmas, legitima la marginación y puede desencadenar dinámicas de segregación y violencia física.

En el plano filosófico, pensadores como Judith Butler han destacado el carácter performativo del lenguaje: las palabras no se limitan a describir la realidad, sino que la constituyen y transforman. En este sentido, el discurso de odio no es meramente ofensivo, sino que produce efectos tangibles sobre la dignidad, la seguridad y la integridad de las personas. Esta concepción ha sido decisiva para fundamentar su regulación en sociedades democráticas, donde la protección de la libertad de expresión debe convivir con el respeto a la dignidad humana.

Desde el punto de vista jurídico, el principal desafío consiste en equilibrar la protección frente al discurso de odio con la salvaguarda del derecho a la libertad de expresión, reconocido como fundamental en la mayoría de las constituciones democráticas y en los tratados internacionales de derechos humanos. La jurisprudencia del Tribunal Europeo de Derechos Humanos ha establecido que la libertad de expresión no ampara manifestaciones que inciten al odio o a la violencia, aunque sí protege opiniones impopulares o provocadoras. De ahí que la delimitación entre discurso crítico y discurso ilícito sea una de las cuestiones más complejas del derecho penal contemporáneo.

El análisis del discurso de odio exige, por tanto, una aproximación interdisciplinaria que considere el contexto comunicativo, la intención del emisor, el impacto en el receptor y las consecuencias sociales. Esta complejidad adquiere especial relevancia en el entorno digital, donde el anonimato, la desinformación y la viralización amplifican los efectos del discurso de odio y dificultan su regulación efectiva.

En este contexto, el auge de los discursos de odio en las sociedades contemporáneas ha supuesto un desafío jurídico de primer orden. En España, su tratamiento penal responde a la necesidad de proteger los valores constitucionales de la dignidad humana, la igualdad y la convivencia pacífica frente a manifestaciones que erosionan la cohesión social y lesionan derechos fundamentales. El delito de odio se configura, así, como una herramienta jurídica compleja que intenta equilibrar la libertad de expresión —núcleo esencial del Estado democrático— con la salvaguarda de los colectivos vulnerables frente a la discriminación y la violencia simbólica.

El reconocimiento jurídico del discurso de odio como categoría autónoma tiene raíces internacionales. Tras la Segunda Guerra Mundial, los instrumentos de protección de los derechos humanos subrayaron la necesidad de limitar aquellas expresiones que promovieran

la intolerancia o la hostilidad entre grupos sociales. El artículo 20.2 del Pacto Internacional de Derechos Civiles y Políticos (ONU, 1966) obliga a los Estados a prohibir por ley “toda apología del odio nacional, racial o religioso que constituya incitación a la discriminación, la hostilidad o la violencia”.

En Europa, el Convenio Europeo de Derechos Humanos (1950), aunque no contiene una disposición específica sobre el discurso de odio, ha sido interpretado por el Tribunal Europeo de Derechos Humanos (TEDH) como habilitante para limitar expresiones que lesionan la dignidad de grupos históricamente discriminados. Este marco se vio reforzado por la Recomendación nº R (97) 20 del Consejo de Europa, que define el discurso de odio como “todas las formas de expresión que difundan, inciten, promuevan o justifiquen el odio racial, la xenofobia, el antisemitismo o cualquier otra forma de odio basada en la intolerancia”.

2.1. Marco normativo: legislación nacional e internacional sobre el discurso de odio

España, como Estado parte de estos instrumentos, ha incorporado progresivamente tales obligaciones. En el plano constitucional, el artículo 20.1 CE garantiza la libertad de expresión, pero el 20.4 CE establece sus límites en el respeto a los derechos reconocidos en el Título I, en particular al honor, a la intimidad y a la dignidad humana. Entendiendo el alto Tribunal en numerosas sentencias que la libertad de expresión no ampara manifestaciones que inciten al odio o a la violencia. Esta circunstancia ha llevado al legislador a tipificar penalmente los comportamientos más lesivos en esta materia.

El principal referente normativo es el artículo 510 del Código Penal, que tras su reforma por la Ley Orgánica 1/2015 amplió su ámbito para incluir tanto la incitación pública al odio, hostilidad, discriminación o violencia, como la producción y difusión de materiales que promuevan tales fines. La conducta típica exige la concurrencia de un dolo específico: el ánimo de incitar o fomentar el odio por motivos discriminatorios (Silva Sánchez, 2022). El precepto protege la dignidad colectiva de los grupos vulnerables, actuando como límite objetivo a expresiones que rebasan la mera opinión ideológica para convertirse en agresión simbólica (Asúa Batarrita, 2019). Asimismo, el artículo 22.4 CP incorpora la agravante genérica de motivación discriminatoria, aplicable a cualquier delito común cuando el móvil se base en la pertenencia de la víctima a un grupo determinado. El artículo constituye el núcleo de la respuesta penal frente a los delitos de odio y las manifestaciones de discurso discriminatorio. Su redacción actual deriva de la profunda reforma operada por la Ley Orgánica 1/2015, que amplió su ámbito subjetivo y objetivo para adaptarlo a los estándares europeos de protección frente a la incitación al odio. El precepto castiga a quienes públicamente fomenten, promuevan o inciten, directa o indirectamente, al odio, hostilidad, discriminación o violencia contra un grupo o persona por motivos racistas, antisemitas u otros referentes a ideología, religión o creencias, situación familiar, pertenencia a una etnia, raza o nación, origen nacional, sexo, orientación o

identidad sexual, género, enfermedad o discapacidad. La amplitud de las conductas típicas responde al mandato constitucional de respeto a la dignidad humana (art. 10 CE) y a la igualdad (art. 14 CE), pero debe interpretarse con cautela para evitar colisionar con la libertad de expresión del artículo 20 CE. En este sentido, el Tribunal Supremo ha reiterado que el tipo penal no sanciona las opiniones, por más repugnantes que sean, sino los actos de incitación que generen un riesgo real de hostilidad social.

La doctrina penal coincide en que el bien jurídico protegido es la dignidad colectiva de los grupos vulnerables y, en última instancia, la convivencia pacífica (Silva Sánchez, 2022; Asúa Batarrita, 2019). No obstante, persiste debate sobre si el tipo requiere una efectiva incitación al odio o basta la mera difusión de mensajes discriminatorios. La Fiscalía General del Estado (2024) sostiene que es necesaria una interpretación restrictiva que preserve la proporcionalidad penal y evite la banalización del tipo. Sin embargo, una parte de la doctrina identifica el bien jurídico protegido no tanto con la dignidad humana y la igualdad (Mir Puig, 2021), sino que otros autores, como Quintero Olivares (2020), doctrinan al orden público como presupuesto de la convivencia plural. Esta dualidad repercute en la interpretación del dolo: no basta con el desprecio o la animadversión hacia un grupo, sino que debe existir una voluntad de incitar a terceros a adoptar actitudes discriminatorias o violentas.

Además de la conducta de incitación, el artículo 510 CP sanciona la producción o distribución de materiales idóneos para provocar tales sentimientos, la negación o justificación de delitos de genocidio y la humillación pública de las víctimas de terrorismo. Se trata, por tanto, de un tipo penal complejo y heterogéneo, que combina delitos de peligro abstracto con otros de resultado simbólico. Esta estructura ha suscitado críticas por su amplitud y por la dificultad de delimitar el dolo específico exigido (Carbonell Mateu, 2020).

El artículo 510 CP representa un instrumento esencial del Derecho penal español para combatir la intolerancia, pero su eficacia depende de una interpretación equilibrada que salvaguarde tanto la dignidad de los colectivos protegidos como el pluralismo democrático. La correcta aplicación del precepto exige, por tanto, una ponderación caso por caso, guiada por los principios de necesidad, proporcionalidad y mínima intervención penal.

No obstante, la aplicación judicial de estos preceptos presenta dificultades interpretativas. En primer lugar, la delimitación entre discurso de odio y libertad de expresión requiere ponderar el contexto, la intención y la potencialidad lesiva del mensaje. El Tribunal Supremo ha sostenido que el tipo penal del artículo 510 CP se refiere a expresiones que trascienden la crítica o la provocación para incitar directamente al rechazo social o a la violencia. En línea con el TEDH, la jurisprudencia española exige que la restricción a la libertad de expresión sea “necesaria en una sociedad democrática”.

El auge de las redes sociales ha multiplicado los casos en los que se cuestiona esa frontera. La instantaneidad y el anonimato de los entornos digitales amplifican la difusión de mensajes potencialmente ilícitos. La Fiscalía de Sala contra los Delitos de Odio y Discriminación ha señalado un incremento constante en las investigaciones relacionadas con discursos en plataformas como X (antiguo Twitter) y Facebook, especialmente por motivos raciales, ideológicos y de orientación sexual (Fiscalía General del Estado, 2024). Sin embargo, el porcentaje de condenas sigue siendo bajo, debido a las dificultades probatorias y al principio de intervención mínima del Derecho penal.

En el plano doctrinal, autores como Carbonell Mateu (2020) y Muñoz Conde (2021) destacan la tensión entre la protección simbólica de los colectivos vulnerables y el riesgo de criminalizar discursos minoritarios o disidentes. Se advierte que un uso expansivo del tipo penal podría afectar la esencia del pluralismo democrático. Por ello, la jurisprudencia insiste en exigir un componente objetivo de incitación y una intencionalidad clara. La Fiscalía General del Estado ha propuesto reforzar la formación de jueces y fiscales en materia de delitos de odio, así como mejorar la cooperación con las plataformas digitales para la identificación de los autores.

En el ámbito nacional, el III Plan de Acción contra los Delitos de Odio (2025–2028), impulsado por el Ministerio del Interior, consolida un marco operativo que refuerza la prevención, la investigación y la atención a las víctimas. El Plan estructura sus líneas estratégicas en torno a la formación de los profesionales del sistema judicial y policial, la mejora de la cooperación institucional y el fortalecimiento de la confianza de las víctimas mediante la reducción de la infradenuncia. Asimismo, incorpora el uso responsable de tecnologías emergentes —incluida la inteligencia artificial— para la detección y el análisis del discurso de odio en entornos digitales, en coordinación con el sector académico y las organizaciones sociales (Ministerio del Interior, 2025).

En este contexto normativo, la aplicación práctica de las disposiciones legales relativas al discurso de odio ha sido objeto de una interpretación constante por parte de los tribunales. El análisis jurisprudencial permite comprender cómo los órganos judiciales —tanto nacionales como europeos— han delimitado los márgenes entre la libertad de expresión y la incitación al odio, ofreciendo criterios esenciales para la correcta aplicación del marco legal.

2.2. Interpretación judicial: jurisprudencia constitucional, penal y europea sobre el discurso de odio

El Tribunal Constitucional ha reafirmado que la libertad de expresión, protegida por el artículo 20 de la Constitución Española, no ampara manifestaciones que inciten a la violencia o discriminación. En sentencias recientes, se ha delimitado el discurso de odio como aquel que lesiona la dignidad humana y pone en riesgo la convivencia democrática.

En la Sentencia del Tribunal Supremo 1234/2018 reafirmó que los delitos de odio no sólo deben entenderse como ataques directos físicos o verbales, sino que también abarcan conductas que fomentan la exclusión social de ciertos colectivos mediante actos discriminatorios sostenidos en el tiempo.

Otra sentencia de gran relevancia es la sentencia del Tribunal Supremo, número 5678/2020, donde se estableció que los discursos de odio, incluso si están enmarcados en contextos de "libertad de expresión", son punibles si se demuestran intenciones discriminatorias o hay un riesgo claro de incitación a la violencia.

Un ejemplo común de incitación al Odio a través de redes sociales, en la vida real se da en las redes sociales, donde personas publican mensajes que promueven el odio hacia ciertos colectivos, como inmigrantes, o individuos de una determinada religión. Estos mensajes, si alcanzan una difusión notable y pueden incentivar a otros a realizar actos violentos, pueden constituir un delito de odio.

Otro ejemplo frecuente se encuentra en el ámbito laboral. Supongamos que un empleador despidе a un empleado alegando su pertenencia a una minoría étnica. Si se demuestra que la motivación del despido está basada en prejuicios racistas, estamos ante un delito de odio según lo estipulado en el Código Penal.

España ha comenzado a integrar herramientas de IA en el sistema judicial. El Ministerio de Justicia y el Consejo General del Poder Judicial han aprobado una Política de uso de la IA en la Administración de Justicia, que busca mejorar la eficiencia mediante soluciones automatizadas para la gestión de casos, análisis de jurisprudencia y redacción de sentencias.

El III Plan de Acción contra los Delitos de Odio (2025–2028), liderado por el Ministerio del Interior, consolida una estrategia integral orientada a la prevención, la investigación y la atención a las víctimas, con especial atención a la mejora de la coordinación interinstitucional y a la reducción de la infradenuncia, uno de los principales obstáculos en la persecución efectiva de estos delitos.

Proyectos como IA4JUSTICE, impulsado en Cataluña, utilizan modelos de lenguaje para asistir en la redacción de resoluciones judiciales, reduciendo significativamente el tiempo de trabajo. Sin embargo, se advierte sobre los riesgos de sesgos algorítmicos y la necesidad de mantener la supervisión humana en todo el proceso.

Además, algunas comunidades autónomas han desarrollado leyes específicas para proteger a colectivos vulnerables, lo que ha generado cierta disparidad en la aplicación de la normativa. Esta fragmentación legal plantea la necesidad de una armonización normativa que garantice una protección equitativa en todo el territorio nacional.

El marco legal internacional sobre el discurso de odio y su regulación mediante inteligencia artificial (IA) se fundamenta en una serie de tratados, recomendaciones y principios desarrollados por organismos como la Organización de las Naciones Unidas (ONU), el Consejo de Europa, la Unión Europea (UE) y la UNESCO, entre otros.

La ONU, a través del Plan de Acción de Rabat, establece criterios para distinguir entre libertad de expresión y discurso que constituye incitación a la discriminación, hostilidad o violencia. Este plan propone una prueba de seis parámetros: contexto, estatus del emisor, intención, contenido, alcance y probabilidad de daño. La ONU enfatiza que combatir el discurso de odio no implica censurar la libertad de expresión, sino prevenir su escalada hacia formas más graves de violencia.

El Consejo de Europa, por su parte, ha adoptado la Recomendación CM/Rec(2022)16, que propone un enfoque integral para combatir el discurso de odio. Distingue entre expresiones que deben ser penalizadas, aquellas que pueden ser reguladas por vía civil o administrativa, y otras que requieren respuestas alternativas como la educación o el contra-discurso.

En cuanto al uso de IA, el Convenio Marco sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho del Consejo de Europa establece principios para garantizar que los sistemas de IA respeten los derechos humanos a lo largo de todo su ciclo de vida. Este tratado, abierto a la adhesión global, busca asegurar la transparencia, la rendición de cuentas y la supervisión ética en el desarrollo y aplicación de IA.

La UNESCO también ha adoptado la Recomendación sobre la Ética de la Inteligencia Artificial, el primer instrumento normativo mundial en esta materia. Este documento aborda riesgos como la vigilancia masiva, la discriminación algorítmica y la falta de transparencia, proponiendo políticas para mitigar estos efectos y promover un uso justo y equitativo de la IA.

Al respecto, la Unión Europea ha integrado el Código de Conducta contra la incitación ilegal al odio en línea dentro de la Ley de Servicios Digitales (DSA). Este código, firmado por plataformas como Facebook, TikTok y YouTube, establece compromisos para eliminar contenidos de odio de forma rápida y proporcional, respetando los derechos fundamentales.

Además, la Asamblea General de la ONU ha adoptado una resolución histórica en 2024 que insta a los Estados a abstenerse de utilizar sistemas de IA que no respeten los derechos humanos, y promueve el desarrollo de marcos normativos inclusivos y seguros para su aplicación.

En síntesis, la jurisprudencia española y europea ha consolidado una interpretación que equilibra la libertad de expresión con la protección frente al discurso de odio, exigiendo siempre la presencia de un riesgo real de incitación a la violencia o la discriminación. Sin embargo, la complejidad del entorno digital y el volumen de información generada en redes

sociales plantean nuevos retos para los operadores jurídicos. En este escenario, la inteligencia artificial emerge como una herramienta estratégica para apoyar la detección y el análisis de estos fenómenos, aunque su incorporación al ámbito jurídico requiere garantías éticas, técnicas y procesales que aseguren el respeto a los derechos fundamentales.

3. Inteligencia artificial y procesamiento del lenguaje natural

Si bien los avances en inteligencia artificial (IA) y procesamiento del lenguaje natural han ampliado las posibilidades de identificar de manera automatizada los discursos de odio en entornos digitales, su aplicación práctica presenta todavía importantes limitaciones. La complejidad semántica del lenguaje, la diversidad cultural y los sesgos inherentes a los datos de entrenamiento evidencian la necesidad de abordar los desafíos técnicos que condicionan la fiabilidad y legitimidad del uso de estas herramientas en contextos jurídicos y sociales.

La IA ha transformado radicalmente la forma en que se aborda el análisis del lenguaje en entornos digitales, especialmente en la detección del discurso de odio. Una de sus ramas más relevantes en este contexto es el procesamiento del lenguaje natural (PLN), que permite a los sistemas comprender, interpretar y generar lenguaje humano de manera automatizada.

El PLN se basa en técnicas de aprendizaje automático (machine learning) y, más recientemente, en aprendizaje profundo (deep learning), que utilizan redes neuronales para identificar patrones complejos en grandes volúmenes de texto. Estas técnicas permiten entrenar modelos con corpus etiquetados que contienen ejemplos de discurso de odio, lenguaje ofensivo, amenazas, insultos y otros tipos de violencia verbal. A partir de estos datos, los sistemas aprenden a clasificar nuevos textos con cierto grado de precisión.

Entre los modelos más utilizados se encuentran los transformers, como BERT, RoBERTa y GPT, que han demostrado una alta capacidad para captar el contexto semántico y sintáctico de las expresiones. Estos modelos pueden analizar no solo palabras aisladas, sino también el significado implícito en frases completas, lo cual es crucial para detectar formas sutiles de odio, como el sarcasmo, la ironía o el doble sentido.

Sin embargo, la aplicación del PLN en la detección del discurso de odio enfrenta varios desafíos. Uno de ellos es la ambigüedad lingüística, que puede llevar a errores de clasificación. Otro es la variabilidad cultural y contextual del lenguaje: una expresión considerada ofensiva en un país puede no tener la misma connotación en otro. Además, el lenguaje evoluciona rápidamente en redes sociales, con el uso de neologismos, emojis, memes y abreviaciones que dificultan la interpretación automática.

Para mejorar la eficacia de estos sistemas, se están desarrollando enfoques híbridos que combinan IA con análisis humano, así como técnicas de explicación algorítmica que permiten entender cómo y por qué un modelo ha clasificado un contenido como discurso de odio. Esto

es especialmente importante en contextos judiciales, donde la transparencia y la trazabilidad de las decisiones automatizadas son fundamentales.

En esta línea, el III Plan de Acción contra los Delitos de Odio (2025–2028) contempla, por primera vez, la utilización de herramientas de inteligencia artificial y análisis de big data con fines preventivos. Estas medidas persiguen mejorar la detección temprana de discursos de odio en redes sociales, reforzar la cooperación entre administraciones y promover la investigación aplicada a la protección de los derechos humanos en el ámbito digital. El Plan destaca la importancia de garantizar la transparencia y la proporcionalidad en el uso de estas tecnologías, promoviendo la colaboración con universidades, centros de investigación y entidades del tercer sector para asegurar un desarrollo ético y socialmente responsable (Ministerio del Interior, 2025).

4. Desafíos técnicos en la detección automatizada

La detección automatizada del discurso de odio mediante inteligencia artificial (IA) enfrenta múltiples desafíos técnicos que limitan su precisión, fiabilidad y aplicabilidad en contextos jurídicos y sociales. Aunque los avances en procesamiento del lenguaje natural (PLN) han mejorado la capacidad de los sistemas para identificar contenido problemático, persisten obstáculos significativos que requieren atención interdisciplinaria.

Uno de los principales retos es la ambigüedad semántica del lenguaje. El discurso de odio puede manifestarse de forma explícita o implícita, utilizando sarcasmo, ironía, dobles sentidos o referencias culturales que escapan a la comprensión literal. Los modelos de IA, aunque entrenados con grandes corpus de datos, tienen dificultades para captar estas sutilezas, lo que puede llevar a falsos positivos (clasificar como odio lo que no lo es) o falsos negativos (no detectar contenido realmente dañino).

Otro desafío es la variabilidad lingüística y cultural. El significado de ciertas expresiones puede variar según el país, la región, el grupo social o el contexto histórico. Por ejemplo, términos que en un entorno son considerados ofensivos pueden ser resignificados positivamente en otro. Esto exige que los sistemas de IA sean sensibles al contexto y adaptables a diferentes comunidades lingüísticas, lo cual es técnicamente complejo y requiere entrenamiento localizado.

Además, el lenguaje evoluciona rápidamente en redes sociales. El uso de emojis, memes, abreviaciones, hashtags y neologismos introduce nuevas formas de comunicación que los modelos tradicionales no siempre reconocen. La actualización constante de los algoritmos es necesaria para mantener su eficacia, lo que implica un esfuerzo continuo de reentrenamiento y validación.

La detección multilingüe también representa un reto. Muchos sistemas están optimizados para el inglés, pero el discurso de odio se expresa en múltiples idiomas y dialectos. La falta de corpus etiquetados en lenguas menos representadas limita la capacidad de los modelos para operar globalmente.

Por otro lado, existe el problema de los sesgos algorítmicos. Los modelos de IA pueden reproducir prejuicios presentes en los datos de entrenamiento, lo que lleva a una discriminación inadvertida contra ciertos grupos. Por ejemplo, pueden ser más propensos a clasificar como discurso de odio expresiones provenientes de minorías lingüísticas o culturales, perpetuando desigualdades.

Los algoritmos deben ser auditables y comprensibles para que sus decisiones puedan ser justificadas ante tribunales. La opacidad de los modelos de aprendizaje profundo plantea un desafío ético y jurídico, ya que dificulta la rendición de cuentas.

En conjunto, estos desafíos técnicos subrayan la necesidad de desarrollar sistemas híbridos que combinen IA con supervisión humana, enfoques contextuales y marcos normativos sólidos que garanticen la equidad, la precisión y el respeto a los derechos fundamentales.

5. Implicaciones legales y éticas

El empleo de sistemas de inteligencia artificial (IA) para la detección del discurso de odio en entornos digitales plantea profundas implicaciones jurídicas y éticas que deben ser abordadas con la máxima rigurosidad. Si bien estas tecnologías constituyen herramientas de gran potencial para la prevención de la violencia digital, su aplicación sin un marco normativo claro y garantista puede generar riesgos sustantivos para los derechos fundamentales reconocidos en los instrumentos internacionales de derechos humanos (UNESCO, 2021).

Desde la perspectiva jurídica, uno de los principales desafíos reside en la validez y admisibilidad de las pruebas generadas por sistemas automatizados. Conforme al artículo 6 del Convenio Europeo para la Protección de los Derechos Humanos y de las Libertades Fundamentales (1950), toda persona tiene derecho a un proceso equitativo, lo que exige que las pruebas sean verificables y que el procedimiento de obtención sea transparente. Sin embargo, los algoritmos de aprendizaje profundo operan como “cajas negras”, lo que dificulta la trazabilidad de las decisiones y la comprensión de los criterios que las fundamentan. Tal opacidad puede comprometer el derecho a la tutela judicial efectiva, reconocido en el artículo 24 de la Constitución Española, así como la presunción de inocencia y la posibilidad de contradicción procesal (Consejo de Europa, 2021).

Un segundo ámbito de preocupación es el de la responsabilidad jurídica derivada del uso de IA. La Propuesta de Reglamento de la Unión Europea por el que se establecen normas armonizadas en materia de inteligencia artificial (AI Act, 2021) introduce un enfoque basado en el riesgo y

exige mecanismos de supervisión humana, pero la atribución de responsabilidad civil o penal en casos de daños ocasionados por sistemas autónomos aún carece de un consenso doctrinal consolidado (Parlamento Europeo y Consejo de la UE, 2021). En este sentido, se plantea la necesidad de establecer marcos de responsabilidad objetiva o compartida que garanticen la reparación efectiva de los daños derivados de errores algorítmicos.

En el plano ético, se plantea el dilema entre la protección frente al discurso de odio y la libertad de expresión. Esta última, reconocida en el artículo 10 del CEDH y en el artículo 19 del Pacto Internacional de Derechos Civiles y Políticos (PIDCP, 1966), constituye un derecho esencial para el debate democrático, aunque no absoluto. La Recomendación General Nº 35 del Comité para la Eliminación de la Discriminación Racial (CERD, 2013) subraya que la lucha contra el discurso de odio debe respetar el principio de proporcionalidad (CERD, 2013). Por ello, los sistemas automatizados de moderación de contenidos deben diseñarse para evitar la censura excesiva y preservar la expresión legítima, garantizando siempre la intervención humana en los casos de ambigüedad.

Asimismo, el principio de no discriminación algorítmica se erige como un pilar ético fundamental. El Consejo de Europa (2021), en su Recomendación CM/Rec(2021)1 sobre el impacto de los sistemas algorítmicos en los derechos humanos, insta a los Estados a realizar auditorías algorítmicas periódicas con el fin de identificar y corregir sesgos que puedan afectar a determinados grupos sociales, culturales o lingüísticos. La equidad, la explicabilidad y la rendición de cuentas deben constituir principios rectores en el diseño y aplicación de estos sistemas (UNESCO, 2021).

Finalmente, la protección de los datos personales representa un componente normativo esencial. La monitorización de redes sociales con fines de detección de discurso de odio implica el tratamiento masivo de información sensible, lo que exige el cumplimiento estricto del Reglamento (UE) 2016/679 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales (RGPD). Dicho reglamento consagra los principios de licitud, transparencia, minimización y limitación de la finalidad (Parlamento Europeo y Consejo de la UE, 2016). Además, impone la obligación de realizar una evaluación de impacto en la protección de datos antes de la implementación de sistemas de IA de alto riesgo (art. 35 RGPD).

En suma, la utilización de inteligencia artificial en la gestión del discurso de odio exige un delicado equilibrio entre innovación tecnológica y respeto a los derechos humanos. La gobernanza algorítmica debe estructurarse sobre los principios de legalidad, proporcionalidad, transparencia, no discriminación y responsabilidad, en consonancia con los estándares internacionales de protección de derechos fundamentales y ética digital (UNESCO, 2021).

6. Propuestas de mejora y recomendaciones

La regulación y detección del discurso de odio mediante Inteligencia Artificial (IA) exige un enfoque multidimensional que combine innovación tecnológica, garantías jurídicas, ética aplicada y participación ciudadana. En primer lugar, resulta esencial fomentar el desarrollo de algoritmos transparentes y auditables que permitan explicar las decisiones automatizadas y someterlas a control independiente. Los sistemas de IA utilizados en la detección del discurso de odio deben garantizar la trazabilidad de sus procesos, de modo que puedan ser comprendidos por los operadores jurídicos, revisados por expertos externos y, en caso necesario, justificados ante los tribunales. La claridad técnica y la documentación de estos modelos son condiciones indispensables para reforzar la confianza institucional y ciudadana, así como para asegurar la rendición de cuentas.

Al mismo tiempo, la automatización nunca debe sustituir el juicio humano. Es necesario establecer mecanismos de supervisión compuestos por comités éticos e interdisciplinarios que acompañen la aplicación de la IA en el ámbito judicial. Estos órganos deberían integrar perfiles técnicos, jurídicos, sociológicos y filosóficos, con el objetivo de revisar las decisiones automatizadas, evaluar casos complejos y garantizar la protección efectiva de los derechos fundamentales. Del mismo modo, la formación especializada de los operadores jurídicos se presenta como una condición imprescindible para un uso responsable de la IA. Jueces, fiscales, abogados y cuerpos policiales deben recibir una capacitación constante en tecnologías digitales, sesgos algorítmicos y derechos digitales, de forma que puedan interpretar con criterio las herramientas tecnológicas y evitar su utilización acrítica o desinformada.

Otro aspecto clave es la armonización normativa y la cooperación internacional. España debe avanzar hacia una convergencia con los estándares promovidos por la ONU, el Consejo de Europa y la Unión Europea, especialmente en lo relativo a la regulación de los sistemas algorítmicos y la protección de datos. La cooperación transfronteriza es esencial para afrontar un fenómeno global como el discurso de odio, compartir buenas prácticas y consolidar marcos comunes de actuación que garanticen la coherencia y la eficacia de las medidas adoptadas. Paralelamente, las políticas públicas deben reforzar la protección de los colectivos históricamente vulnerables —personas migrantes, LGTBI+, minorías religiosas o personas con discapacidad—, mediante la mejora de los canales de denuncia, la atención psicológica y jurídica a las víctimas y la promoción de campañas de sensibilización orientadas a la empatía y la inclusión social.

Asimismo, más allá de la censura o la eliminación de contenidos, se debe apostar por la promoción del contra-discurso y la alfabetización digital. Fomentar narrativas inclusivas que promuevan la diversidad y el respeto constituye una estrategia preventiva más eficaz y democrática frente al odio. La educación digital, implementada desde etapas tempranas y

extendida a toda la ciudadanía, permitirá reconocer, denunciar y contrarrestar de manera informada los discursos de odio en línea. Finalmente, toda herramienta de IA aplicada a este ámbito debe someterse a una evaluación continua basada en indicadores de eficacia, equidad y respeto a los derechos humanos. Los mecanismos de retroalimentación con usuarios, víctimas y profesionales son imprescindibles para detectar errores, sesgos o efectos no deseados, garantizando así una mejora constante de los sistemas y un uso más justo, transparente y socialmente responsable de la inteligencia artificial.

7. Enfoque interdisciplinar

La complejidad del discurso de odio en entornos digitales exige un abordaje interdisciplinar que articule los ámbitos tecnológico, social, jurídico y ético. La interacción entre estos campos permite comprender no solo la naturaleza del fenómeno, sino también las posibilidades y límites del uso de la Inteligencia Artificial (IA) en su prevención y control. A continuación, se desarrollan las tres dimensiones fundamentales de esta propuesta.

7.1. Dimensión tecnológica

El avance de la inteligencia artificial aplicada al análisis del lenguaje ha transformado la forma en que se detectan y evalúan los discursos de odio en redes sociales. Entre las herramientas más relevantes se encuentran los modelos basados en procesamiento del lenguaje natural (PLN), como BERT, RoBERTa y su adaptación específica para contenidos ofensivos, HateBERT. Estos modelos emplean redes neuronales profundas que aprenden a identificar patrones semánticos y contextuales a partir de grandes corpus de datos etiquetados, como el Hate Speech Dataset o el proyecto MODERATE, financiado por la Comisión Europea. En el contexto español, el proyecto IA4JUSTICE constituye una referencia clave. Impulsado por el Consejo General del Poder Judicial, integra sistemas de análisis automatizado para la gestión de resoluciones judiciales y la detección de sesgos en decisiones penales. La evaluación de la precisión y fiabilidad de estos modelos se realiza mediante métricas como la F1-score, el recall y la precisión ponderada, que permiten medir su capacidad para detectar correctamente expresiones de odio sin generar falsos positivos. Aun así, estos sistemas enfrentan limitaciones importantes: la ambigüedad lingüística, los sesgos algorítmicos y la dependencia de datos representativos. Por ello, se recomienda la implementación de mecanismos de auditoría y revisión humana permanente que garanticen la transparencia y la trazabilidad de los resultados.

7.2. Dimensión social

El discurso de odio en redes sociales no afecta por igual a todos los grupos sociales. Diversos estudios demuestran que las mujeres, las personas LGTBIQ+, las minorías étnicas y religiosas, los migrantes y las personas con discapacidad son los principales objetivos de ataques discriminatorios. Este tipo de violencia digital reproduce y amplifica desigualdades

estructurales preexistentes, generando un impacto psicológico, simbólico y material que trasciende el ámbito virtual. Asimismo, la brecha digital agrava estas desigualdades: los colectivos más vulnerables suelen tener menor acceso a mecanismos de denuncia y a herramientas de protección tecnológica. La pobreza digital, entendida como la falta de competencias y recursos para defenderse de la violencia en línea, se convierte así en un nuevo factor de exclusión. Desde esta perspectiva, el combate contra el discurso de odio debe enmarcarse en una estrategia de justicia social digital, que combine la regulación penal con políticas educativas, campañas de alfabetización mediática y programas de empoderamiento digital dirigidos a grupos en riesgo. La IA, en este sentido, no debe entenderse solo como una herramienta de control, sino como un instrumento de inclusión y prevención.

7.3. Dimensión ética y de derechos humanos

El desarrollo y aplicación de la inteligencia artificial en la detección del discurso de odio debe regirse por principios éticos que aseguren el respeto a los derechos fundamentales. En el ámbito internacional, documentos como la Recomendación sobre la Ética de la Inteligencia Artificial de la UNESCO (2021) y el Convenio Marco del Consejo de Europa sobre IA, Derechos Humanos, Democracia y Estado de Derecho (2024) establecen estándares para garantizar la transparencia, la rendición de cuentas y la no discriminación algorítmica. En el marco europeo, el Reglamento (UE) 2024/1689, conocido como *Ley de Inteligencia Artificial* (Unión Europea representa un hito regulatorio, al clasificar los sistemas de IA según su nivel de riesgo y exigir mecanismos de supervisión ética y social. Su correcta aplicación resulta esencial para asegurar que la IA se utilice en consonancia con los valores democráticos y la dignidad humana. Esta perspectiva ética se alinea con los Objetivos de Desarrollo Sostenible de la Agenda 2030, especialmente el ODS 10 (Reducción de las desigualdades) y el ODS 16 (Paz, justicia e instituciones sólidas). Ambos objetivos subrayan la necesidad de fortalecer instituciones inclusivas y de promover sociedades que garanticen la igualdad y la justicia a través de la tecnología. En última instancia, el reto no es solo técnico, sino ético y político: garantizar que las herramientas de inteligencia artificial sirvan a la justicia social y a la defensa activa de los derechos humanos. La IA debe concebirse como un medio al servicio de la humanidad, orientado a proteger la dignidad y la diversidad frente a las nuevas formas de violencia simbólica que emergen en el espacio digital.

8. Conclusión

La creciente presencia del discurso de odio en las redes sociales constituye uno de los desafíos más complejos para las sociedades democráticas contemporáneas. Este fenómeno, que adopta múltiples formas —desde insultos explícitos hasta narrativas sutiles de exclusión—, impacta de manera directa en la cohesión social, la dignidad humana y la seguridad de los colectivos más vulnerables. En este escenario, la Inteligencia Artificial (IA) se presenta como una herramienta

de gran potencial para detectar, analizar y contrarrestar estas expresiones, aunque su aplicación plantea interrogantes técnicos, jurídicos y éticos que no pueden ser obviados.

A lo largo del trabajo se ha mostrado que, si bien los avances en procesamiento del lenguaje natural han permitido desarrollar sistemas cada vez más precisos para identificar patrones de odio, persisten limitaciones significativas. La ambigüedad del lenguaje, la variabilidad cultural, la rápida evolución de las formas de comunicación digital y los sesgos algorítmicos constituyen obstáculos que requieren respuestas interdisciplinarias y una supervisión humana constante.

Desde el punto de vista jurídico, el marco normativo español —centrado en el artículo 510 del Código Penal— proporciona una base sólida para la sanción del discurso de odio, pero necesita adaptarse a los retos del entorno digital y a las nuevas exigencias de la justicia algorítmica. La jurisprudencia nacional e internacional ha avanzado en la delimitación de los límites entre la libertad de expresión y la incitación al odio; sin embargo, aún queda un camino por recorrer para lograr un equilibrio estable que garantice la protección efectiva de los derechos fundamentales.

Del mismo modo, el uso de la IA en el sistema judicial español, aunque todavía incipiente, evidencia un potencial transformador. Iniciativas como el proyecto IA4JUSTICE y las políticas impulsadas por el Consejo General del Poder Judicial apuntan hacia una modernización de la justicia, pero su desarrollo debe ir acompañado de garantías éticas, transparencia y formación especializada que aseguren un uso legítimo y responsable de estas tecnologías.

Las propuestas formuladas en este trabajo, que abarcan desde el diseño de algoritmos auditables hasta la promoción del contra-discurso y la alfabetización digital, destacan la necesidad de una estrategia integral que combine tecnología, derecho, educación y participación ciudadana. La lucha contra el discurso de odio no puede sustentarse únicamente en la censura o en la respuesta penal, sino que debe incorporar medidas preventivas, pedagógicas y comunitarias orientadas a construir una cultura digital basada en el respeto, la inclusión y el diálogo.

La reciente aprobación del III Plan de Acción contra los Delitos de Odio (2025–2028) evidencia el compromiso del Estado español con un enfoque integral que combina prevención, protección, educación y tecnología. Al incorporar medidas de sensibilización, apoyo a las víctimas y uso ético de la inteligencia artificial, el Plan refuerza la dimensión interdisciplinar y colaborativa necesaria para abordar el discurso de odio desde una perspectiva democrática y de derechos humanos (Ministerio del Interior, 2025).

En última instancia, la inteligencia artificial puede convertirse en una aliada decisiva para la defensa de los derechos humanos en el entorno digital, siempre que su implementación esté guiada por principios democráticos, supervisión ética y un compromiso firme con la justicia

social. España, al igual que otras naciones, se encuentra ante una encrucijada: aprovechar el potencial transformador de la tecnología sin renunciar a los valores que sostienen el Estado de derecho. El futuro de la regulación del discurso de odio dependerá, en definitiva, de nuestra capacidad colectiva para equilibrar innovación y responsabilidad, eficacia y equidad, automatización y humanidad.

Bibliografía

ASÚA BATARRITA, M., 2019. *Discurso de odio y libertad de expresión: un análisis desde el Derecho penal*. Revista de Derecho Penal y Criminología, 22, pp. 45–68.

CARBONELL MATEU, J., 2020. *Derecho Penal. Parte General*. 5.ª ed. Valencia: Tirant lo Blanch.

CASELLI, T., BASILE, V., MITROVIĆ, J. y GRANITZER, M., 2021. *HateBERT: Retraining BERT for Abusive Language Detection in English*. En: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 17–25. Online: Association for Computational Linguistics. Disponible en: <https://doi.org/10.18653/v1/2021.woah-1.3>

CENTRO DE ESTUDIOS POLÍTICOS Y CONSTITUCIONALES (CEPC), 2024. *Libertad de expresión, odio y redes sociales*. Madrid: CEPC. Recuperado de: <https://www.cepc.gob.es/publicaciones/libertad-expresion-odio-redes>

CERD (Comité para la Eliminación de la Discriminación Racial), 2013. *Recomendación general nº 35: Lucha contra el discurso de odio racista*. Naciones Unidas.

CONSEJO DE EUROPA, 1950. *Convenio Europeo para la Protección de los Derechos Humanos y de las Libertades Fundamentales*. Roma: Consejo de Europa.

CONSEJO DE EUROPA, 2021. *Recomendación CM/Rec(2021)1 sobre el impacto de los sistemas algorítmicos en los derechos humanos*. Estrasburgo: Consejo de Europa.

CONSEJO DE EUROPA, 2024. *Convenio Marco sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho*. Estrasburgo: Consejo de Europa.

CONSEJO GENERAL DEL PODER JUDICIAL (CGPJ), 2023. *Proyecto IA4JUSTICE: Inteligencia Artificial al servicio de la justicia*. Madrid: CGPJ.

FISCALÍA GENERAL DEL ESTADO, 2024. *Memoria de la Fiscalía General del Estado 2024*. Madrid: Ministerio de Justicia.

GORWA, R., BINNS, R. y KATZENBACH, C., 2020. *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*. Big Data & Society, 7(1). Disponible en: <https://doi.org/10.1177/2053951719897945>

HATE SPEECH DATASET, 2020. *Open dataset for multilingual hate speech detection*. Recuperado de: <https://github.com/hate-speech-dataset>

MACAVANEY, S., YAO, H.-R., YANG, E., RUSSELL, K., GOHARIAN, N. y FRIEDER, O., 2019. *Hate speech detection: Challenges and solutions*. PLOS ONE, 14(8), e0221152. Disponible en: <https://doi.org/10.1371/journal.pone.0221152>

MINISTERIO DEL INTERIOR, 2024. *Delitos de odio en España: estado de la cuestión*. Madrid: Gobierno de España. Recuperado de: <https://www.interior.gob.es/opencms/es/servicios-al-ciudadano/delitos-de-odio/>

MINISTERIO DEL INTERIOR, 2025. *III Plan de Acción de Lucha contra los Delitos de Odio (2025–2028)*. Madrid: Gobierno de España. Disponible en: https://www.interior.gob.es/opencms/pdf/archivos-y-documentacion/documentacion-y-publicaciones/publicaciones-descargables/seguridad-ciudadana/3_plan-accion-delitos-odio_126250364_pdfWEB.pdf

MODERATE PROJECT, 2023. *Monitoring of Online Discourse to Eradicate Hate Speech*. Comisión Europea, Programa Horizon Europe.

MUÑOZ CONDE, F., 2021. *El discurso de odio del artículo 510.1.a) CP*. Revista de Derecho Penal y Criminología, 26, pp. 143–168.

NACIONES UNIDAS, 2015. *Agenda 2030 para el Desarrollo Sostenible*. Nueva York: ONU.

ORGANIZACIÓN DE LAS NACIONES UNIDAS (ONU), 2019. *Defining hate speech: United Nations Strategy and Plan of Action on Hate Speech*. Nueva York: ONU. Recuperado de: <https://www.un.org/en/hate-speech/understanding-hate-speech>

PACTO INTERNACIONAL DE DERECHOS CIVILES Y POLÍTICOS, 1966. *Naciones Unidas*. Nueva York: ONU.

PARLAMENTO EUROPEO Y CONSEJO DE LA UNIÓN EUROPEA, 2016. *Reglamento (UE) 2016/679 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales (RGPD)*. Diario Oficial de la Unión Europea.

PARLAMENTO EUROPEO Y CONSEJO DE LA UNIÓN EUROPEA, 2021. *Propuesta de Reglamento por el que se establecen normas armonizadas en materia de inteligencia artificial (AI Act)*. Bruselas.

SILVA SÁNCHEZ, J. M., 2022. *Los delitos de odio en el Derecho penal español: problemas interpretativos*. Revista Electrónica de Ciencia Penal y Criminología, 24(3), pp. 1–32.

TRIBUNAL SUPREMO (SALA SEGUNDA), 2018. *Sentencia 396/2018, de 9 de febrero (RJ 2018/938)*. Madrid: Centro de Documentación Judicial.

UNESCO, 2021. *Recomendación sobre la Ética de la Inteligencia Artificial*. París: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. Recuperado de: <https://unesdoc.unesco.org/ark:/48223/pf0000379920>

UNITED NATIONS, 2024. *Addressing hate speech in the digital era: global report*. Nueva York: United Nations Publications.

UNIÓN EUROPEA, 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 (Ley de Inteligencia Artificial o “AI Act”)*. Diario Oficial de la Unión Europea, L 202, 12 jul. 2024, pp. 1–108. Disponible en: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

Sergio Pérez González y David San Martín Segura. *Ciando* por los Derechos Humanos

Sergio Pérez González

Profesor Titular de Derecho Penal, Universidad de La Rioja

David San Martín Segura

Profesor Permanente Laboral de Derecho Administrativo, Universidad de La Rioja

[ACCEDE A LA GRABACIÓN](#)

Resumen

El Colectivo de Inteligencia Artificial e Inteligencia Social (CIAIS) está compuesto por personal investigador de la Universidad de La Rioja (actualmente 44 personas de 16 áreas y 7 departamentos diferentes), cuyos ámbitos de investigación van desde la informática hasta la filología o la ética, incluyendo diversas ramas jurídicas. Lo que nos vertebra como Colectivo es el análisis de las transformaciones provocadas por los recientes desarrollos de las tecnologías informáticas, tanto en los niveles del poder como en las distintas estructuras y grupos sociales. Las siglas CIAIS pretenden evocar el término náutico «ciar», que significa remar hacia atrás: creemos que el método de aproximación científica requerido en la actualidad implica tomar perspectiva para propiciar reflexiones de calidad, sin dejar de prestar atención al día a día.

Desde abril de 2023 hemos desarrollado múltiples actividades, con un método de trabajo que evidencia la emergencia de nexos entre inteligencia artificial y derechos humanos. Esta actividad se ordena en torno a tres objetivos.

Primero. Desplegar hacia la sociedad, mediante acciones divulgativas, una invitación a comprender y asumir los efectos -positivos y negativos- de los avances de las tecnologías informáticas.

Segundo. Facilitar, a diversas ramas del saber, el conocimiento y acceso a ciertas herramientas de inteligencia artificial (o asociadas) que no suelen estar a su alcance, fundamentalmente por su elevado coste.

Tercero. Construir nuevas líneas de investigación a partir de la integración de conocimientos provenientes de los diversos campos científicos del personal del CIAIS. Se trata de ir más allá de lo meramente interdisciplinar, con un objetivo nuevo y original: integrar conocimientos económicos, jurídicos, sociológicos o medioambientales (entre otros) en la investigación en informática.

En esta presentación defenderemos que en nuestros días resulta evidente que las tecnologías no pueden desarrollarse sin asumir el postulado de que en las sociedades actuales lo tecnológico nunca es neutral.

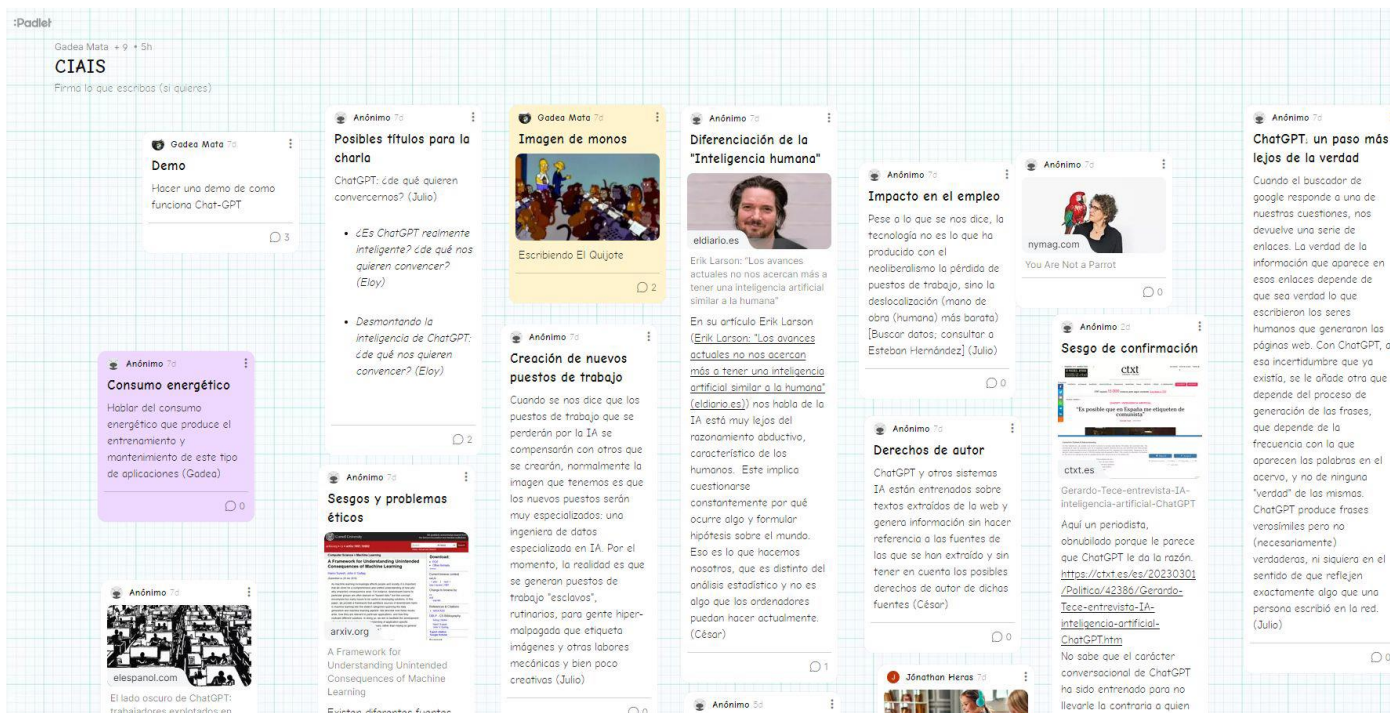
Introducción

El Colectivo de Inteligencia Artificial e Inteligencia Social (CIAIS) está compuesto por personal investigador de la Universidad de La Rioja de muy distintas áreas de conocimiento. En concreto, los miembros del CIAIS pertenecen a 7 departamentos diferentes del total de 12 en los que se divide la Universidad de La Rioja (por orden alfabético: Agricultura y Alimentación, Ciencias Humanas, Derecho, Economía y Empresa, Filologías Hispánicas y Clásicas, Filologías Modernas y Matemáticas y Computación), y a 16 áreas de conocimiento (Análisis Matemático, Ciencia de la Computación e Inteligencia Artificial, Derecho Administrativo, Derecho Constitucional, Derecho Penal, Didáctica de las Ciencias Experimentales, Didáctica de las Matemáticas, Filología Inglesa, Filosofía Moral, Historia Medieval, Lengua Española, Lenguajes y Sistemas Informáticos, Literatura Española, Organización de Empresas, Producción Vegetal y Tecnología de los Alimentos).

Tan variado aparataje conceptual converge en la atención por la transversalidad tecnológica del mundo. Y es que el acrónimo CIAIS se hace polisémico y, convenientemente tildado, evoca la segunda persona del plural del presente de indicativo de «ciar», término náutico que significa remar hacia atrás: la intervención tecnológica del mundo hace necesario ahormar epistemologías diversas para interpretar la realidad con fiabilidad descriptiva y confluir en un impulso conjunto de toda la nave. Esa promiscuidad metodológica aspira a contrarrestar la inercia del imperativo tecnológico, a cuestionar la instrumentalidad autorreferente de la tecnicidad.

Sobre esta base, la finalidad global del CIAIS es la de fomentar la investigación básica y aplicada en los campos de la inteligencia artificial, las tecnologías informáticas, la ciencia de datos y las ciencias humanas y sociales, poniendo énfasis en las interacciones entre todas estas disciplinas. No se trata, por tanto, de mera colaboración interdisciplinar, sino de integración de perspectivas en los fundamentos del conocimiento de la tecnología, así como en proyectos aplicados específicos. Para ello, la actividad de CIAIS procura siempre la interacción para la detección de usos sospechosos de la IA, como se visibiliza en el *padlet* elaborado por todos sus miembros y que sirvió como palanca iniciática del proyecto¹⁴⁰:

¹⁴⁰ Disponible en <https://padlet.com/gadeamata/ciais-c6190yvvxmycx1ol>



Este leitmotiv global se desglosa en tres objetivos generales:

El primer objetivo general es el de aportar una mirada crítica hacia los nuevos desarrollos en tecnologías de la información desde una perspectiva humanística. Existe una tendencia a que la investigación y el desarrollo en informática e inteligencia artificial esté guiado por fines específicos, sin atender efectos éticos o sociales de más amplio espectro. La experticia jurídica, lingüística, sociológica o filosófica permite que la investigación en tecnología se desarrolle desde fundamentos epistemológicos mucho más amplios e inclusivos. Según la clásica disección weberiana de las formas de racionalidad, se trataría de modular la desnuda instrumentalidad formal de la tecno-ciencia con aportaciones de racionalidad sustantiva de cariz ético, jurídico, sociológico o político.

El segundo objetivo general es el de transferir las herramientas informáticas, matemáticas, estadísticas y de inteligencia artificial a aplicaciones en torno a los saberes generados desde otras disciplinas, señaladamente las sociales y humanas. Suele suceder que son las herramientas informáticas las que condicionan las posibilidades de la investigación, como si se tratase de una evolución tecnológica paralela que el personal investigador debe aprovechar al paso. Este efecto se ha multiplicado con la aparición de los desarrollos en Inteligencia Artificial, de modo que los métodos y objetivos de las ciencias quedan cada vez más modulados por sus instrumentos. Incluso en campos gnoseológicos propios de las «ciencias duras» (*α-operatorias*), sus formas de modelización características se están viendo desplazadas por las técnicas de la ciencia de datos vehiculadas en sistemas de IA. Esas interferencias epistémicas no son menores en los campos de las ciencias humanas, sociales y jurídicas. Así, resulta conveniente la colaboración directa entre especialistas de los ámbitos tecnológicos y humanísticos para

revertir la preponderancia en la comunicación y facilitar la diferenciación entre lo necesario y lo decorativo en las labores de investigación.

El tercer objetivo general es el de promover la divulgación de calidad en la frontera de las disciplinas implicadas. En un doble sentido: por un lado, es necesario que el personal científico e investigador valore las novedades tecnológicas en un horizonte de eficacia complejo, que incorpore la evaluación de efectos utilitarios en la propia disciplina, pero que también evalúe el impacto ético, social o legal. Por otro lado, es necesario fomentar entre la ciudadanía una cultura de la tecnología de juicio equilibrado y contextualizado, que valore las condiciones concretas en las que surge y se implementa, sin recurrir a posiciones primarias tecno-entusiastas o tecno-fóbicas. Por ello es fundamental realizar actividades que permitan abordar estas inquietudes con un enfoque accesible al público en general.

Con estos antecedentes, en esta comunicación pretendemos defender que las tecnologías, de modo autónomo, no pueden desarrollarse de manera neutral o aun azarosa, sino que, necesariamente, se inscriben en tendencias marcadas por intereses parciales o aun en confluencias ideológicas. La tesis basal, así, pasa por considerar que los sistemas de inteligencia artificial que condicionan o aun determinan decisiones en muy distintos ámbitos resultan difícilmente controlables desde criterios de racionalidad ética, social o jurídica. La incidencia de esta ausencia de control en la garantía de los derechos humanos es un riesgo cada vez más inasumible.

Para exponer esta tesis, proponemos tres pasos: en primer lugar, exponemos una base teórica común a las líneas de investigación diversas que se plantean desde CIAIS; un suelo teórico que explique, de modo transversal, que la tecnologización progresiva y la codificación en lenguaje informático plantea una mediación del mundo que exige, constantemente, un salto desde el lenguaje natural con pérdida de significado y, por tanto, con pérdida de control (§2). En segundo lugar, planteamos situaciones concretas en las que esta pérdida de control genera, en ámbitos muy diversos, pérdida de racionalidad social; se trata de una relación de situaciones que han sido abordadas en actividades divulgativas concretas por parte del CIAIS y que, por desgracia, son solo una evidencia parcial de la pérdida progresiva de racionalidad social (§3). En último lugar, se proyectan afecciones concretas que estas situaciones pueden generar en el derecho como disciplina y en los derechos humanos, en particular, como libertades y garantías inherentes a toda persona, independientemente del derecho objetivo en el que se despliegan (§4).

2. Mediación tecnológica del mundo y pérdida de racionalidad

2.1. El problema de la tecnología

Que la técnica no puede pensarse solo como instrumento o medio para ciertos fines es una intuición filosófica clásica que la eclosión de la IA parece certificar a cada paso innovador. En La

pregunta por la técnica, Heidegger afrontaba esta cuestión con el bagaje fenomenológico correspondiente: «la técnica no es lo mismo que la esencia de la técnica»¹⁴¹. La noción griega clásica de *poiesis* ayuda a descriptar estas tribulaciones heideggerianas: la técnica, en su esencia, es un modo de *desocultamiento*, una forma de revelación por reacción de la observación. Sin embargo, la técnica moderna no revela el mundo de la misma manera que el arte o la producción artesanal (*techne*), sino que transforma las relaciones en estrategias de control y utilidad. Escribe el filósofo alemán: «la técnica moderna es una provocación que pone ante la naturaleza la exigencia de suministrar energía que como tal pueda ser extraída y almacenada»¹⁴².

Dicho de otro modo, el problema de la tecnología como técnica refinada, como técnica desarrollada desde recetas científicas, es que no muestra lo que ya está, como en un desvelar aséptico. El mundo se muestra como recurso a disposición, pero no como presencia preconfigurada. La consecuencia es que la esencia de la tecnología tiende a ocultar otras formas posibles de revelación del mundo. Heidegger advierte de la ilusión más peligrosa: «Cuando del peor modo estamos abandonados a la esencia de la técnica es cuando la consideramos como algo neutral, porque esta representación (...) nos hace completamente ciegos para la esencia de la técnica»¹⁴³. El problema principal, entonces, no es el daño que puedan causar las máquinas en las rutinas humanas (que nos sustituyan en el hacer), sino el modo en que la tecnología nos relaciona conceptualmente con el mundo, mostrándolo de un modo entre otros posibles que se mantienen ocultos (que nos sustituyan epistemológicamente). El riesgo esencial es que quedemos cegados ante otras formas de verdad y de presencia.

Pero la perspectiva del alemán, como la nuestra, no es por ello tecnófoba. «Somos capaces de experimentar lo técnico en su limitación»¹⁴⁴, dice Heidegger, de modo que hay resquicios para desbordar la tecnología del quicio utilitario en el que se desarrolla principalmente. Cabe un pensar *poietico*, poético. Así, el problema de la tecnología no es que existan máquinas e inteligencia artificial, sino que estas monopolicen el modo de revelar el mundo, que debería mantenerse siempre abierto a nuevas sensibilidades.

¹⁴¹ Heidegger, M. «La pregunta por la técnica». En: *Conferencias y artículos*. Barcelona: Ediciones del Serbal, 1994, p. 10

¹⁴² *Ibidem*, p. 17.

¹⁴³ *Ibidem*, p. 9.

¹⁴⁴ *Ibidem*, p. 9.

En España, la obra de Ortega y Gasset puso el problema de la tecnología en el centro del debate filosófico ya a finales de los años 30 ¹⁴⁵. Ortega comparte buena parte de la crítica heideggeriana¹⁴⁶: la técnica no es simplemente un conjunto de medios para fines ni un añadido externo a la vida humana. Para él, lo técnico pertenece a la estructura más íntima de lo humano. Y es que el humano se hace a sí mismo creando mundo y creando posibilidades. La técnica, en este sentido, es constitutiva del ser humano, no un instrumento que lo circunda accesoriamente. Es cierto que la tecnología es «el esfuerzo para ahorrar esfuerzo»¹⁴⁷, pero en un sentido amplio, no como medio para satisfacer necesidades inmediatas, sino como modo de crear un entorno vital adecuado a un proyecto. Ortega explica que el humano no vive simplemente para resolver carencias, sino para realizar posibilidades, para inventar nuevas formas de vida: «La técnica es la reforma de la naturaleza»¹⁴⁸. Y, en esta clave, quizá la mirada que, transportada a la actualidad, resulte más lúcida es la que concibe el desarrollo acelerado de la tecnología como un peligro que desajusta la sociedad, incapaz de dotar de sentido a tanta posibilidad abierta en tan poco tiempo. Ortega llegó a esta conclusión mediado el siglo XX¹⁴⁹. La irrupción de los sistemas de IA nos reafirma en su denuncia. Y es que, si la tecnología se desarrolla para la realización de la vida humana, el riesgo es que la vida quede desorientada por demasiados planes incomprensibles.

Confluyen las críticas clásicas de la técnica en el concepto de *cosmotécnica* de Yuk Hui¹⁵⁰, según el cual cada cultura desarrolla una forma particular de entender y aplicar la tecnología, influenciada por su cosmovisión y contexto histórico. Es por ello ficticia la noción de una única «tecnología universal», desprendida de todo contexto, como si el progreso tecnológico solo pudiera ser uno, transversal y homogeneizador de toda cultura. Hui aboga por la «tecnodiversidad», de modo que las posibilidades de interacción en un entorno cultural determinen su tecnología, y no al revés.

2.2. La pérdida de racionalidad social

La ausencia de neutralidad de la tecnología cristaliza en los tiempos de procesamiento de datos a gran escala en un problema más concreto. La recodificación del pensamiento, el deseo y, en

¹⁴⁵ Ortega y Gasset, J. «Meditación de la técnica». En: *Obras completas*. Vol. V. Madrid: Revista de Occidente, 1964.

¹⁴⁶ O al revés: Heidegger fue alumno de Ortega en su estancia en Friburgo en 1929.

¹⁴⁷ Ortega y Gasset, J. «Meditación de la técnica». En: *Obras completas*. Vol. V. Madrid: Revista de Occidente, 1964, p. 331.

¹⁴⁸ *Ibidem*, p. 324.

¹⁴⁹ Ortega dice: «(...) la técnica, al aparecer por un lado como capacidad, en principio ilimitada, hace que al hombre, puesto a vivir de fe en la técnica y sólo en ella, se le vacíe la vida (...) De puro llena de posibilidades, la técnica es mera forma hueca —como la lógica más formalista—; es incapaz de determinar el contenido de la vida» (*Ibidem*, p.366).

¹⁵⁰ Hui, Y. *La cuestión de la tecnología en China: Un ensayo sobre cosmotécnica*. Buenos Aires: Caja Negra Editora, 2016.

general, la comunicación humana en lenguaje computacional es el salto de traducción que permite replicar la actividad y el horizonte vital del ser humano en sistemas de IA. Se trata, entonces, de analizar la traslación de lenguaje natural a lenguaje informático.

Entre dos lenguajes siempre hay desajustes de traducción. Sin embargo, este problema, en la mayoría de las ocasiones, ha sido enfrentado como un problema de traducción entre dos lenguajes naturales y, digamos con la tradición fenomenológica, lenguajes mundanos, vivenciales; esto es, entre dos lenguajes surgidos en comunidades lingüísticas distintas, pero con mundos de experiencia cotidiana relativamente parangonables. La cuestión, por tanto, cambia su naturaleza cuando el lenguaje que pretende ser traducido no surge desde el mismo mundo de experiencia que el lenguaje al que quiere ser traducido. En parte, se trata del problema de la *inconmensurabilidad* que afrontaba Thomas Kuhn en su *Estructura de las revoluciones científicas*¹⁵¹. El concepto de inconmensurabilidad surge primero en el ámbito de los paradigmas científicos. Es cierto que, en este marco, Kuhn lo aborda de una manera relativamente vaga, pero con el tiempo desarrollaría esa primera intuición proyectándola sobre las teorías científicas: en el entorno de una revolución científica, dos teorías -una a un lado y otra al otro del corte revolucionario- son difícilmente traducibles; esto es, son difícilmente significables en uno de los dos códigos descriptivos o en un tercero. Hay implicada en esta tesis de Kuhn una teoría del conocimiento e incluso una teoría del ser relativamente reconocible. Y es que las distintas comunidades científicas observan el mundo conforme a sus aparatos tecnológicos y conceptuales, de tal modo que las supuestas descripciones objetivas de mundo no solo serán distintas, sino que serán incomparables entre sí, porque las agrupaciones de similitudes y diferencias de objetos será distinta, al menos en parte (el ejemplo típico es el del concepto *planeta*, que modifica su extensión tras la revolución copernicana). Kuhn introduciría el concepto de interpretación para sustituir al de traducción entre teorías inconmensurables. La interpretación, a diferencia de la traducción, incorpora glosas a los enunciados teóricos originales, ya que solo con estos añadidos pueden colmarse los huecos entre explicaciones de mundos. Para ello, indica Kuhn, el intérprete debe conocer la teoría que va a interpretar, de modo que pueda usar alguno -o la mayoría- de sus elementos en la teoría posterior.

Si traducimos, por ejemplo, del inglés al español, podemos encontrar diferencias en los mundos de experiencia de los grupos de hablantes, pero estas serán fácilmente subsanables con herramientas de interpretación coyunturales. Sin embargo, ya no puede hablarse de traducción si intentamos comprender un enunciado de física cuántica con el aparato conceptual de la física

¹⁵¹ Kuhn, T. S. *La estructura de las revoluciones científicas*. 4ª ed. México: Fondo de Cultura Económica (Breviarios), 2017, pp. 101-103.

clásica, por acudir al ejemplo más socorrido de *ciencia dura*. En ese marco, ya ha debido desarrollarse un ejercicio de interpretación complejo (puede atisbarse una revolución científica entre ambos aparatos conceptuales).

Si trasladamos el problema de la inconmensurabilidad al ámbito que aquí nos interesa en relación con el lenguaje informático, los desajustes de traducción se verán sublimados. Y es que el asunto adquiere una complejidad mayúscula si los enunciados forman parte de aparatos conceptuales distintos no solo por razón de distinta comunidad lingüística de origen o porque el instrumental de descripción del mundo haya evolucionado y mutado en parte, sino porque esos aparatos conceptuales están, en su mismo fundamento, contruidos sobre claves distintas (ya no es que el telescopio o el microscopio electrónico abran una *puerta*, sino que se observa y describe el mundo, *ab initio*, con utensilios y tecnologías diversas). Esto sucede al trasladar significados del lenguaje natural al lenguaje informático lógico-formal-matemático por otro. Este es el problema de traducir instrucciones mundanas a lenguaje informático. Pero tampoco es un asunto nuevo: ya a principios del siglo XX el matemático Gottlob Frege, aunque anterior a la informática, distinguía entre *referencia* y *sentido* a propósito de la posibilidad de transformar el lenguaje natural en proposiciones lógicas, de modo que la referencia sería el objeto o entidad concreta a la que se refiere la expresión, mientras que el sentido es el modo o la manera en que esa referencia se presenta o se entiende. Y en esa balanza hizo equilibrios Wittgenstein: el del *Tractatus logico-philosophicus*¹⁵² sostenía que el lenguaje funciona como una representación lógica del mundo, donde las proposiciones serían *figuras* de los hechos¹⁵³; y el de las *Investigaciones filosóficas*¹⁵⁴ problematizaría esa visión mostrando la complejidad del uso natural del lenguaje, lo que anticiparía los límites de traducir el lenguaje ordinario en términos formales¹⁵⁵.

En el ámbito de una sola disciplina -en su evolución, que es lo que preocupaba a Kuhn-, estas controversias tenían sentido en relación con teorías o enunciados que competían por representar más certeramente la realidad (con algún instrumento distinto). En esa lógica, la prueba de la experiencia es la que designa la teoría ganadora. Asimismo, resulta crucial una posición fuerte en el plano epistemológico o aun ontológico, ya que la experiencia parece someter a las teorías a un *juicio de realidad*. Sin embargo, cuando el referente de descripción

¹⁵² Wittgenstein, L. *Tractatus logico-philosophicus*. Madrid: Tecnos, 2017.

¹⁵³ Esa es la base de su «teoría figurativa del lenguaje» recogida en la tesis 4.01: «La proposición es una figura de la realidad. La proposición es un modelo de la realidad tal como pensamos que es».

¹⁵⁴ Wittgenstein, L. *Investigaciones filosóficas*. Madrid: Editorial Trotta, 2021.

¹⁵⁵ «En cuanto al término significado, éste puede ser explicado —para muchos casos de su utilización, aunque no para todos— como sigue: el significado de un término es su uso en el lenguaje» (*Ibidem*, p. 311).

no es una realidad «ahí afuera», sino, precisamente, un primer enunciado que ha de ser trasladado a otro lenguaje deviene menos importante la adscripción ontológica.

Así las cosas, la cuestión de fondo radica en que lenguaje natural y lenguaje informático no son traducibles a uno de los dos o a un tercer lenguaje. Este era el debate sobre el reduccionismo científico en el siglo XIX al que condujeron los postulados más radicales del positivismo lógico de Carnap o Neurath (lo biológico puede explicarse en lo químico y lo químico, a su vez, en lo físico, que sería la base lingüística única para explicar el mundo, ya que es la disciplina más explicable, a su vez, en enunciados lógico-matemáticos)¹⁵⁶. Dejado atrás el *sueño positivista*, debe enfrentarse que lenguaje natural y lenguaje informático se han desarrollado a partir de experiencias de mundo distintas, distinción que trae causa, a su vez, de distintas mediaciones instrumentales con la realidad. En el caso del lenguaje informático, los pulsos eléctricos mediante los que se codifica determinan la generación de un sistema formal de enunciados mucho más finos que los del lenguaje natural; estos, por el contrario, a partir de la aprehensión humana, conforman un sistema de enunciados que, en las ciencias sociales, pero también -de modo distinto- en muchas ciencias naturales, está repleto de solapamientos, ambigüedades, polisemias, homonimias, sinonimias, paronimias, etc., necesitado de un contexto extralingüístico que lo haga funcionar con cierta eficacia.

En ese salto de «granularidad» que relaciona los lenguajes de modo muy distinto con lo real se enmarca el problema de inconmensurabilidad y, por tanto, de necesidad de interpretación en la traslación de enunciados. En la práctica, el proceso de interpretación puede intentar controlarse en una secuencia que pasa por lenguajes intermedios (diagramas de flujo y otras formas de especificación formal). Con todo, siempre hay una barrera invisible que separa el dominio del mensaje desde el lado natural (poco a poco reduciéndose a comandos lógicos) hasta transformarse en codificación formal pura.

John Searle lo explicó con éxito en su famoso argumento de la habitación china, en el que cuestiona que la mera traducción entre lenguaje natural y lenguaje computacional pueda captar significado o comprensión. La sintaxis, esto es, las reglas formales de manipulación simbólica, no basta para generar semántica (entendimiento). Del ejemplo se infiere que la simulación externa (generar un *output* con apariencia de entendimiento) no implica comprensión profunda; porque la comprensión profunda, en cierto modo, reconfigura no solo la comunicación directa, sino -aunque imperceptiblemente- el horizonte de vida completo¹⁵⁷.

¹⁵⁶ Ayer, A. J. (comp.). *El positivismo lógico*. México: Fondo de Cultura Económica, 1981.

¹⁵⁷ Searle, J. R. *Mentes, cerebros y ciencia*. Madrid: Editorial Cátedra, 2001.

También Chomsky, en un sentido parecido, refiere la diferencia entre *competencia* (el sistema ideal de conocimiento lingüístico) y *actuación* (el uso concreto, con errores, limitaciones de memoria, distracciones...). La imperfección humana, sus límites de actuación, reconfiguran las competencias. Y eso es difícil, o imposible, de replicar computacionalmente. Luciano Floridi, en fin, explica y valora sentencioso los efectos de esta transcripción computacional: «Una de las cuestiones más espinosas y en cierto modo estimulantes se refiere al modo en que las tecnologías digitales traducen, simplifican y a veces empobrecen el significado original del lenguaje natural, reduciéndolo a conjuntos de datos procesables»¹⁵⁸.

Hay, por tanto, sistemáticamente, un cierto margen de incertidumbre en cualquier codificación informática. Si partimos de un lenguaje natural que ofrece instrucciones básicas para su codificación en lenguaje informático, las posibilidades de desarrollo desborden sistemáticamente las de las instrucciones. Digamos, por usar la expresión de Kuhn, que el *traductor* a lenguaje informático es, en realidad, un *intérprete* que ha de *glosar* las instrucciones, de modo que un enunciado en lenguaje natural siempre puede trasladarse, sin incorrecciones formales, a varios enunciados informáticos. No es banal, por cierto, que quien interprete y glose suela ser un profesional informático, que conoce también el lenguaje natural (a diferencia del jurista, economista o profesional de la medicina, que raramente conoce el lenguaje informático). Pero todas estas disquisiciones adquieren una dimensión inédita cuando el lenguaje informático generado conforma un sistema de inteligencia artificial, dotado, a su vez, de capacidad de aprendizaje automático. Ya no solo la interpretación en el tránsito de lenguajes aleja al humano del resultado de la traducción, sino que el ulterior proceso dinámico de detección y organización de patrones extraídos de datos masivos desenchaja al humano de las decisiones en las que va a contextualizarse su vida, que cada vez quedarán más lejos de su sensibilidad.

Las regularidades que construye la máquina a partir de modelos probabilísticos proporcionan al humano respuestas, predicciones, clasificaciones y, en último término, verdades que resultan coherentes con la información que recibió el sistema algorítmico en su entrenamiento. Sin embargo, esas verdades están generadas en procesos en los que el humano cada vez participa menos. Participan datos asignados computacionalmente a humanos. Participa la comunicación generada por sistemas de IA desde referencias primigenias del sentir humano, pero cada vez peor *traducidas*, cada vez más desactualizadas, más desenfocadas, cada vez más basadas en deseos incrustados, en glosas de interpretación garabateadas por presunciones estadísticas utilitarias. Hay, en definitiva, una intensificación del distanciamiento con el lenguaje natural, de

¹⁵⁸ Traducción propia de: Floridi, L., *Filosofia dell'informazione*, Milano: Raffaello Cortina Editore, 2024, p. 139.

modo que el problema clásico sobre la técnica se agudiza hasta límites insospechados. Y estas decisiones de vida se despliegan en muy diversos ámbitos. En CIAIS, se han identificado algunos.

3. Efectos de la pérdida de racionalidad social y divulgación del CIAIS

En los últimos años, la inteligencia artificial se ha encauzado en el lema del imperativo tecnológico como un progreso inevitable que, en su propia lógica de avances utilitarios, proyecta una cierta imagen de autonomía y neutralidad. Y con este aval, los ámbitos afectados son muchos:

En el ámbito de la salud, los sistemas de inteligencia artificial se desarrollan como asistentes virtuales de pacientes y apoyan a los profesionales sanitarios en la elaboración de diagnósticos y propuesta de tratamientos. En la educación, la inteligencia artificial se utiliza en plataformas que marcan el ritmo y estilo de aprendizaje del estudiantado. En el ámbito laboral, la IA ayuda a seleccionar y evaluar personal mediante el análisis de competencias y desempeños, así como automatiza tareas repetitivas. En economía, la inteligencia artificial se emplea para crear modelos predictivos que anticipan cambios en mercados y optimiza la gestión de recursos. Sin embargo, una mirada crítica sobre el modo en que se diseñan y despliegan los sistemas de IA visibiliza los intereses espurios implicados en su desarrollo e implementación. Los efectos en muy distintos ámbitos son indicios de ello¹⁵⁹, y en CIAIS los abordamos de manera específica:

Hay en el uso de la IA una tensión constante entre facilitación cortoplacista y daño invisibilizado. En buena medida, esta tensión reproduce la del interés individual, específico, inmediato y el coste social que, en el largo plazo, coloca al colectivo en una situación más indefensa¹⁶⁰. Las actividades del CIAIS han procurado visibilizar esta tensión casi siempre soterrada. En la salud, por ejemplo, proyectos de simplificación de textos médicos se programan como un modo de facilitar la comprensión de información sensible, de reducir barreras comunicativas y de favorecer aparentemente la autonomía del paciente; pero esconden el riesgo de reducir la complejidad clínica a estadísticas escogidas y hacer tender la toma de decisiones a parámetros de selección sesgados e interesados¹⁶¹. En el ámbito educativo, experiencias innovadoras que

¹⁵⁹ «Tecnología y sociedad: el caso de la Inteligencia Artificial», celebrada el 1 de diciembre de 2023 en el IES Celso Díaz (Arnedo, La Rioja). Charla de Julio Rubio.

¹⁶⁰ Curso de verano «Implicaciones sociales de la IA», celebrado del 15 al 19 de septiembre de 2025 en Arnedo (La Rioja). Más información en: <https://www.unirioja.es/actividades/xxxi-curso-de-verano-ciudad-de-arnedo-implicaciones-sociales-de-la-inteligencia-artificial/>

¹⁶¹ «Simplificación de textos médicos con procesamiento del lenguaje: el proyecto CLARA-MeD», celebrada el 23 de mayo de 2023 en la Universidad de La Rioja. Charla a cargo de Leonardo Campillos, dentro del Seminario de Informática Mirian Andrés. Más información en: <https://seminariomirianandres.unirioja.es/2023/05/23/simplificacion-de-textos-medicos-con-procesamiento-del-lenguaje-el-proyecto-clara-med/>

ponen a estudiantado y docentes frente a sistemas generativos pueden convertirse en ejercicios críticos y formativos, pero solo si se plantean como una confrontación controlada y no como una delegación acrítica en la máquina¹⁶². La interacción con modelos generativos puede ser una herramienta pedagógica si se la encuadra en proyectos reflexivos¹⁶³. Pero la riqueza del razonamiento humano, con sus dudas, contradicciones y contextos, no puede reducirse a respuestas probables de un sistema que solo *simula* comprensión¹⁶⁴.

La necesidad de reducir el alejamiento de *lo humano* es, en algunos ámbitos, especialmente evidente. En el contexto del Trabajo Social, por ejemplo, mediaciones de sistemas de IA condicionan con criterios poco claros intervenciones que implican sensibilidades difícilmente codificables¹⁶⁵. El impacto de estas tecnologías en las violencias estructurales es igualmente decisivo. El cruce entre *big data*, internet y violencia de género revela cómo los algoritmos no solo reproducen estereotipos y sesgos, sino que en ocasiones los amplifican¹⁶⁶. De ahí la necesidad de aproximaciones feministas que examinen las economías de plataforma, cada vez más fiscalizadoras de nuestras rutinas, y muestren cómo se inscriben en relaciones de dominación¹⁶⁷. Y La cultura digital, atravesada por algoritmos de recomendación y posicionamiento, muestra con claridad las consecuencias de esta lógica¹⁶⁸. Los sistemas traducen conductas humanas en patrones estadísticos y devuelven preferencias que parecen

¹⁶² Proyecto de innovación docente «Mi avatar corrige a ChatGPT», concedido el 9 de enero de 2024 en la Universidad de La Rioja. Participantes: Delia del Pilar Gavela (Coordinadora), Sara Gómez, Paula Jiménez, Clara Jiménez, Mirari San Martín, Vanessa Alvear y Gadea Mata.

¹⁶³ «Aplicando la IA generativa con espíritu crítico en el contexto universitario», Jornada de innovación educativa celebrada el 27 de junio de 2024 en la Universidad de La Rioja. Con intervenciones de Thais Ruiz de Alda y Delia Gavela.

¹⁶⁴ «Desmontando la inteligencia de ChatGPT: ¿de qué nos quieren convencer?», conferencia de Julio Rubio y mesa redonda con la participación de Olaya Fernández, Sergio Pérez, Gadea Mata y Juan Manuel Agulles, moderada por Delia del Pilar Gavela, celebrada el 12 de abril de 2023 en el Ateneo Riojano. Más información en: <https://ateneoriojano.com/events/charla-desmontando-la-inteligencia-de-chatgpt-de-que-nos-quieren-convencer/>

¹⁶⁵ «Inteligencia Artificial, Ciencia Ciudadana y Trabajo Social», reunión científica celebrada el 15 de noviembre de 2023 en la Universidad de La Rioja. Intervención de Gadea Mata en la mesa redonda Tecnologías aplicadas a los servicios sociales y al Trabajo Social. Más información en: <https://www.unirioja.es/actividades/reunion-cientifica-inteligencia-artificial-ciencia-ciudadana-y-trabajo-social/>

¹⁶⁶ «Violencia de Género. Inteligencia Artificial, Big Data e Internet», curso de verano celebrado el 4 de julio de 2023 en la Universidad de La Rioja. Intervención de Gadea Mata en el Taller de Violencias Digitales (I). Más información en: <https://www.unirioja.es/actividades/cursos-de-verano-violencia-de-genero-inteligencia-artificial-big-data-e-internet/>

¹⁶⁷ «Una mirada feminista a la economía de plataforma y la inteligencia artificial», Jornada celebrada el 7 de marzo de 2024 en la Universidad de La Rioja. Presentación de Julio Rubio y David San Martín. Ponencias de Débora Ávila, Núria Soto, César Domínguez y Sergio Pérez. Más información en: <https://www.unirioja.es/actividades/jornada-cia-is-una-mirada-feminista-a-la-economia-de-plataforma-y-la-inteligencia-artificial/>

¹⁶⁸ «Entre algoritmos y audiencias. Una perspectiva interdisciplinar sobre el posicionamiento SEO», sesión celebrada el 18 de noviembre de 2024 en la Universidad de La Rioja. Intervenciones de Ignacio Marco y Sandra Ramírez. Más información en: <https://www.unirioja.es/actividades/sesion-entre-algoritmos-y-audiencias-una-perspectiva-interdisciplinar-sobre-el-posicionamiento-seo/>

naturales, pero que son el producto de una simplificación que privilegia algunas rentabilidades¹⁶⁹.

Tampoco puede ignorarse la dimensión ecológica: el consumo de agua y energía que implica la infraestructura de IA contradice cualquier relato de innovación limpia o sostenible. El entusiasmo tecnológico olvida con demasiada facilidad que estas máquinas requieren de recursos naturales limitados y que su expansión multiplica la presión sobre ecosistemas ya frágiles¹⁷⁰.

Frente a este panorama, es indispensable una ética de la inteligencia artificial traducible en prácticas institucionales y sociales capaces de orientar el desarrollo tecnológico hacia objetivos sociales¹⁷¹. La clave está en asumir que no se trata de aceptar el progreso técnico como un destino inevitable, sino de integrarlo en una racionalidad social y política que priorice derechos, justicia y sostenibilidad. Si al pasar del lenguaje natural al lenguaje informático se pierde significado, señaladamente significado social, es necesario diseñar salvaguardas institucionales y sociales que compensen esa pérdida. De lo contrario, los sistemas seguirán operando con una lógica que erosiona la deliberación democrática.

El conflicto ético abre un desafío particular para la recodificación computacional de normas jurídicas. La incorporación de sistemas automáticos en ámbitos públicos convierte los algoritmos en auténticas normas de facto, pues se institucionalizan prácticas que afectan derechos¹⁷². La opacidad de estos procesos plantea nuevos problemas para el Derecho público¹⁷³ y exige pensar en cómo garantizar transparencia y control democrático¹⁷⁴. En este

¹⁶⁹ «Dime qué escuchas y te diré en quién te convierten: análisis crítico de la influencia de las plataformas digitales», celebrada el 26 de junio de 2025 en el Seminario Mirian Andrés (CCT). Charla de Yoseba Peña.

¹⁷⁰ «Inteligencia por medios artificiales y su consumo de recursos naturales: lo que pocas personas imaginan», celebrada el 17 de junio de 2025 en la Universidad de La Rioja. Charlas de Vicente Santiago Marco Mancebón y Félix Lanas Mangado. Más información en: <https://www.unirioja.es/actividades/inteligencia-por-medios-artificiales-y-su-consumo-de-recursos-naturales-lo-que-pocas-personas-imaginan/>

¹⁷¹ Conferencia de Txetxu Ausín («Ética para convivir con la inteligencia artificial y la ciencia de datos») en la Jornada de apertura de la IV edición del Máster en Ciencia de Datos y Aprendizaje Automático; y ponencias de Pilar Agustín, David San Martín, Olaya Fernández y Emma Juaneda. Más información en: <https://www.unirioja.es/actividades/conferencia-inaugural-4a-edicion-del-master-universitario-en-ciencia-de-datos-y-aprendizaje-automatico/>

¹⁷² «Conceptos artificiales: justicia, inteligencia, política», celebrada el 19 de noviembre de 2024 en la Universidad de La Rioja. Proyección de la película Justicia Artificial y posterior debate. Más información en: <https://www.unirioja.es/actividades/conceptos-artificiales-justicia-inteligencia-politica/>

¹⁷³ «Normas para ‘software’, ‘software’ para normas», sesión celebrada el 17 de marzo de 2025 en la Universidad de La Rioja. Intervenciones de Juan José Olarte, David San Martín y Diego Prieto. Más información en: <https://www.unirioja.es/actividades/sesion-normas-para-software-software-para-normas/>

¹⁷⁴ «En busca de la transparencia del algoritmo. Nuevas problemáticas para el Derecho Público», celebrada el 25 de abril de 2023 en la Universidad de La Rioja. Charla a cargo de Sergio Capellán, dentro del Seminario de Informática Mirian Andrés. Más información en: <https://seminariomirianandres.unirioja.es/2023/04/25/en-busca-de-la-transparencia-del-algoritmo-nuevas-problematicas-para-el-derecho-publico/>

contexto, 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 inteligencia artificial se presenta como una oportunidad, pero también como un terreno en disputa, ya que la regulación tecnológica no puede ser, tan solo, un marco de legitimación para una práctica deshumanizadora¹⁷⁵.

El reto es calibrar la doble normatividad que, según nos advierte Simondon, exhibe todo objeto técnico¹⁷⁶. El producto tecnológico opera según una normatividad interna que responde a su tecnicidad misma y es independiente de los usos que externamente se proyecten sobre el objeto: cualquier uso estará condicionado por esas prescripciones inherentes al objeto que son el resultado de su desarrollo evolutivo, de su afinamiento técnico. Por otro, todo producto tecnológico tiene una existencia social (es un objeto socialmente inmerso) y, por lo tanto, está atravesado por normatividades externas que condicionan en distinto grado su realidad técnica. La discusión sobre el desarrollo tecnológico y la aprehensión de sus efectos ha de moverse en esos dos planos normativos.

La cuestión, en definitiva, no es cómo convivir con máquinas inteligentes, sino cómo construir un marco democrático en el que la inteligencia artificial sea una herramienta subordinada a criterios sociales, jurídicos y éticos. Sin ese horizonte, las tecnologías seguirán funcionando como dispositivos de poder opacos. Y este marco de control racional complejo, como intentamos visibilizar en CIAIS, solo es posible, al margen de normativas de amplio espectro discursivo, con intervenciones específicas en ámbitos y contextos concretos en los que desvelar intereses espurios y construir resistencias.

4. Afectación a los derechos humanos

Como apuntábamos, el problema general de la tecnología como mediación universal tiene también un impacto directo en el derecho. En concreto, la agudización de esta cuestión por el uso indiscriminado de sistemas de IA se proyecta en un doble sentido: en primer lugar, el derecho debe asimilar en sus respectivas dogmáticas efectos y aun daños provocados por sistemas autónomos; en segundo lugar, el derecho afronta el dilema de desarrollarse a partir de decisiones generadas por sistemas autónomos.

¹⁷⁵ «Derecho y sistemas de IA: Implicaciones para el Derecho español del reglamento en materia de Inteligencia Artificial de la Unión Europea», celebrada en 2024 en la Universidad de La Rioja. Con participación de profesorado de varias ramas jurídicas. Más información en: <https://www.unirioja.es/actividades/derecho-y-sistemas-de-ia-implicaciones-para-el-derecho-espanol-del-reglamento-en-materia-de-inteligencia-artificial-de-la-union-europea/>

¹⁷⁶ Simondon, G. *El modo de existencia de los objetos técnicos*. Buenos Aires: Prometeo Libros, 2013 (1ª ed. en francés, 1954).

4.1. El derecho frente a sistemas autónomos

En el primer sentido apuntado, el derecho ha dispuesto soluciones transversales, señaladamente 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 inteligencia artificial. Su objeto general es establecer un marco jurídico común para el desarrollo, uso y comercialización de sistemas de inteligencia artificial en la Unión Europea, garantizando que sean seguros, transparentes y respetuosos con los derechos fundamentales. Se basa en un enfoque proporcional al nivel de riesgo que presenta cada sistema de IA. En España, se encuentra en tramitación el Proyecto de Ley para el Buen Uso y la Gobernanza de la Inteligencia Artificial. Su finalidad es adaptar la normativa nacional al Reglamento europeo, promoviendo un uso ético y responsable de la IA. Asimismo, el Real Decreto 729/2023, de 22 de agosto, creó formalmente la Agencia Española de Supervisión de la Inteligencia Artificial (AESIA), dedicada específicamente a la supervisión y promoción del uso ético y confiable de la inteligencia artificial en España, en coherencia con el marco europeo.

Sin embargo, más allá de esta transversalidad, los efectos de la incertidumbre incorporada a las instrucciones en lenguaje jurídico natural o a las especificaciones técnicas proveídas desde ciertas posiciones de responsabilidad son muy variados. Sin duda, una respuesta jurídica general no puede absorber el impacto social y personal del uso de sistemas de IA, que escapa a las prohibiciones o límites de trazo grueso. La ubicuidad e impregnación de la IA requiere esfuerzos adaptados desde cada ámbito jurídico concreto, ya que las exigencias dogmáticas de cada grupo normativo varían: qué grado de relación causalidad e imputación personal puede exigirse en derecho penal ante un daño producido por IA; qué grado de responsabilidad objetiva y qué reparto de consecuencias indemnizatorias son exigibles en derecho civil según qué determinaciones automáticas se hayan dado; qué grado de riesgo puede permitirse en derecho mercantil para que las empresas operen automáticamente en un determinado mercado de valores, etc.

El derecho, por tanto, debe desarrollar nuevas pautas de control y racionalidad adaptadas del uso de IA en su heterogeneidad de dogmáticas. Y para ello los riesgos deben precisarse con rigor y con crítica radical, esto es, tomando en consideración el problema elemental de la implementación tecnológica desreferenciada. Como indica el Reglamento (UE) 2021/1689, se ha de prestar especial atención «a las maneras en que los modelos pueden dar lugar a sesgos dañinos y discriminación que entrañan riesgos para las personas, las comunidades o las sociedades, a la facilitación de la desinformación o el menoscabo de la intimidad, que suponen una amenaza para los valores democráticos y los derechos humanos» (Considerando 110). Pero

esa consideración que toma por referencia la posible afectación de derechos humanos¹⁷⁷ debe protegerse en flancos jurídicos diversos. La tutela efectiva frente a tales desafíos solo podrá ofrecerse con una identificación del peligro en una clave crítica integral y no solo coyuntural. De lo expuesto, las afectaciones a derechos humanos que proponemos en esta comunicación como fruto del trabajo y las perspectivas compartidos en CIAIS son las siguientes:

En relación con el derecho de participación política, el riesgo de uso de la IA en el sentido expuesto es que la esfera pública quede limitada por una lógica estadística tendente a excluir disidencias. En ese contexto, el derecho de participación política se vería erosionado por mecanismos que filtran la participación en foros públicos a través de algoritmos privados. Que la participación política, en sus diferentes niveles, sea un ejercicio deliberativo incorpora resituaciones de sensibilidad que no pueden reducirse a un proceso de retroalimentación de datos. Las decisiones políticas quedarían influenciadas por sistemas de recomendación, segmentación de audiencias o incluso generación automatizada de discursos y consignas que, basadas en correlaciones extraídas de comportamientos previos, marcarían la agenda pública. El riesgo de representación política por la estadística, por agrupaciones de datos a las que ni el sujeto concreto conoce pertenecer, es, por ello, reseñable.

Este desplazamiento tendría consecuencias profundas, ya que la deliberación pública reduciría la complejidad del pensamiento político a patrones utilitarios y parciales. En la medida en que la mediación algorítmica traduce el lenguaje político en métricas de rendimiento o impacto, el derecho a participar en la vida pública perdería su sustancia dialógica.

En relación con el derecho a la libertad de expresión, la gestión algorítmica redefine los circuitos en los que circula y se visibiliza la comunicación. Lo que se expresa ya no depende tanto de la voluntad del hablante como de los parámetros de pertinencia, *viralidad* o adecuación definidos por sistemas de clasificación automatizados. Los discursos son jerarquizados, amplificados o silenciados según métricas de eficiencia comunicativa, ajena al hablante y su autoridad en otros parámetros. Así, el derecho formal de expresarse libremente cohabita con la restricción material del alcance de lo dicho. Asimismo, la progresiva modulación del lenguaje natural por agentes artificiales adaptados a cada pequeña interacción (mensajería instantánea, email, chat de consulta...) introduce una censura estructural que asienta incluso una sintaxis maquinal. La expresión pierde su densidad relacionada directamente con la sensibilidad humana. En este escenario, la producción de comunicación aumenta ostensiblemente, pero el significado

¹⁷⁷ Véase Consejo de Europa. *Recommendation CM/Rec (2020)1 on the human rights impacts of algorithmic systems*. Estrasburgo: Council of Europe, 2020.

renovador o aperturista se estrecha. Y no por órdenes verticales desde fuera, sino por determinación de la tecnología que creíamos instrumental.¹⁷⁸

En general, los entornos de relación digital son contextos de sociabilidad intensamente mediada y sometida a vigilancias métricas que aspiran a valorizar las interacciones. Las redes sociales hiperbolizan las dinámicas de *imitación, oposición e innovación* que Gabriel Tarde intuyó anticipadamente como elementos consustanciales de nuestra urbanidad contemporánea¹⁷⁹. El ágora digital exacerba las influencias imitativas, las actitudes de rechazo, los esfuerzos por exhibir originalidad. Todo ello condicionado por las reglas de visibilidad algorítmicamente definidas. La socialización se desarrolla al ritmo de los parámetros de medición del deseo y de la creencia que cimentan aquellos entornos de interacción digital.

En relación con el derecho al trabajo y a unas condiciones laborales justas, la automatización del trabajo humano bajo la lógica de la eficiencia algorítmica condiciona una tutela real del derecho subjetivo. Muchas tomas de decisión laboral, como la contratación, la asignación de tareas o la evaluación de rendimiento, se llevan a cabo por modelos probabilísticos que convierten al sujeto trabajador en datos interpretados en métricas de productividad. El derecho al trabajo, cuya gestación histórica ha erosionado, por vía de la participación social, el monopolio de los criterios productivistas, afronta ahora un riesgo muy cierto. Las plataformas digitales, bajo la apariencia de neutralidad técnica y, por tanto, sin sesgos ideológicos aparentes que puedan identificarse con interés de parte, determinan horarios, frecuencias de entrega y evaluaciones mediante una monitorización que robotiza al humano, que proyecta su horizonte vital en un flujo de datos.

Así, la remuneración adecuada por el trabajo prestado, el descanso apropiado o la conveniencia de la estabilidad ya no surgen de una negociación entre partes que dialogan en lenguaje natural y están expuestas a la presión social o a las preponderancias políticas, sino que se *cientifican* en parámetros de eficacia contruidos sobre resultados económicos pasados frente a los que la resistencia laboral aparece como irracional. El sujeto trabajador queda atrapado en un complejo decisonal virtual que pone en cuarentena los derechos adquiridos.¹⁸⁰

4.2. Derecho desarrollado en sistemas de IA

¹⁷⁸ Véase Zuboff, S. *The Age of Surveillance Capitalism*. New York: PublicAffairs, 2019; también Floridi, L. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford: Oxford University Press, 2014.

¹⁷⁹ Tarde, G. *Les lois de l'imitation. Étude sociologique*, Paris: Félix Alcan éditeur, 1890, p. 97.

¹⁸⁰ El informe de la Agencia de los Derechos Fundamentales de la Unión Europea (FRA) (2021), *Trabajadores de plataformas digitales y derechos fundamentales*. Luxemburgo: Oficina de Publicaciones de la Unión Europea, analiza el impacto de la automatización y la gestión algorítmica sobre derechos como la igualdad, la remuneración justa y la negociación colectiva.

En el segundo sentido apuntado -que el derecho se desarrolle, en normativas o decisiones administrativas y judiciales, mediante sistemas autónomos- el riesgo radica en la replicación de la dualidad referida por Frege, sentido y referencia. El remedo jurídico, entonces, sería el de norma y disposición, de modo que la norma incorporaría el sentido y el significado; mientras que la disposición haría las funciones de referencia, de signo, de expresión formal. El salto de la referencia al sentido es difícilmente computerizable. Y así sucedería, por ejemplo, en relación con el derecho a la salud, que quedaría comprometido como prestación pública. No solo por la brecha digital que generaría, de suyo, una limitación del acceso, sino por los sesgos incorporados al conocimiento y tratamiento de las patologías. Una medicina algorítmica clasificaría grupos de patologías o aun de pacientes como poco relevantes si no se implementan controles de racionalidad éticos. Así, la crítica a la medicina moderna, a la producción de conocimiento médico orientado por intereses farmacéuticos, tomaría un sentido renovado si los criterios de investigación y desarrollo privilegian y automatizan las variables economicistas.

Cuando las tecnologías algorítmicas se emplean para adoptar decisiones jurídicas formales, el Derecho se ve sacudido en sus formas de «razonamiento» y en el modo de manifestarse en cuanto sistema normativo y lingüístico. Esas afecciones serán de distinto grado según las tecnologías aplicadas (codificación de reglas estáticas o bien sistemas con capacidad de aprendizaje autónomo) y en razón del carácter de la decisión automatizada (administrativa resolutoria o de trámite, judicial de unos u otros órdenes jurisdiccionales). Pero hay un cierto grado de afectación inherente que apunta a forma en que la norma es ahora expresada al verse encapsulada en software, lo que suscita problemas de cognoscibilidad de su operativa y de acceso de la ciudadanía a su diseño y concreta codificación. En Derecho público esa automatización tensa necesariamente las categorías, principios y garantías sedimentadas para la sujeción del poder público. Afecta, inevitablemente, a la «lengua de los derechos»¹⁸¹.

Este problema se revela incluso en la aplicación de tecnologías relativamente sencillas, como expresa el conocido «caso Bosco». Tal es el nombre de un software adoptado por la Administración General del Estado para la evaluación automatizada de las solicitudes de percepción del llamado «bono social eléctrico» previsto para consumidores considerados vulnerables ¹⁸². La controversia jurídica de largo recorrido sobre el software, primero administrativa y luego judicial, se suscitó por la pretensión de una Fundación (Civio) dedicada a la observación del ejercicio del poder público de acceder a las características y contenido de

¹⁸¹ Tomamos la expresión del discurso que García de Enterría ofreció en su recepción como académico de la RAE el 24 de octubre de 1994. García de Enterría, E. *La lengua de los derechos. La formación del Derecho Público europeo tras la Revolución Francesa*, Madrid: Alianza, 1994.

¹⁸² <https://www.miteco.gob.es/es/energia/pobreza-energetica/pe-001.html>

dicho software, incluido su código fuente¹⁸³. El motivo no era baladí: el sistema venía arrojando salidas que, en algunos casos, no respondían a la norma «analógica» (un Real Decreto) que regula los criterios de concesión del bono social¹⁸⁴. La Administración General del Estado, el Consejo de Transparencia y Buen Gobierno, el Juzgado Central de lo Contencioso Administrativo y la Audiencia Nacional rechazaron el pretendido derecho de Cívico a acceder al código fuente del software como información pública en manos de la Administración¹⁸⁵. Lo hicieron bajo argumentos de seguridad (riesgos de explotación de vulnerabilidades «de día cero» que pusieran en jaque las infraestructuras digitales del Estado), pero también netamente jurídicos: afirmaron que Bosco era objeto de derechos de propiedad intelectual titularidad de la propia Administración, lo que permitiría excluir el acceso a su código. Ha sido el Tribunal Supremo quien, en casación, ha corregido el sentido del fallo y ha exigido la entrega del código fuente a la Fundación apelante¹⁸⁶. Lo ha hecho, empero, bajo un argumento menos rupturista de lo que pudiera parecer. Sigue el Tribunal afirmando la sujeción de Bosco a derechos de propiedad intelectual, si bien estos deben ceder, en el caso, frente al derecho de acceso a la información pública en una suerte de ponderación de intereses en disputa.

Más allá de sus particularidades, el caso ha hecho explícitos algunos interrogantes sobre el uso de tecnologías algorítmicas por el poder público: ¿qué carácter ha de atribuirse a esas formas de software? ¿es posible afirmar derechos de propiedad intelectual sobre él? ¿cómo afecta el trabajo de codificación informática (que, a menudo, es desarrollado por sujetos privados) a principios elementales del Derecho público? ¿de qué marco debemos dotarnos para juridificar esa clase de software con incidencia decisoria? ¿en qué casos el uso de una tecnología algorítmica ha de considerarse realmente decisorio o meramente auxiliar de una decisión humana?¹⁸⁷

¹⁸³ <https://civio.es/tag/bono-social/>

¹⁸⁴ Real Decreto 897/2017, de 6 de octubre, por el que se regula la figura del consumidor vulnerable, el bono social y otras medidas de protección para los consumidores domésticos de energía eléctrica.

¹⁸⁵ Respectivamente: desestimación de la solicitud de acceso a la información pública formulada por Cívico por parte de la Administración General del Estado, por silencio administrativo; Resolución del Consejo de Transparencia y Buen Gobierno 701/2018, de 18 de febrero de 2019; Sentencia del Juzgado Central de lo Contencioso Administrativo (nº8) 143/2021, de 30 de diciembre de 2021; Sentencia de la Audiencia Nacional (sala de lo contencioso-administrativo) de 30 de abril de 2024 (recurso nº 51/2022).

¹⁸⁶ Sentencia del Tribunal Supremo (sala tercera) 1119/2025, de 11 de septiembre de 2025.

¹⁸⁷ Esta última pregunta ha estado en el centro del debate jurídico en torno Bosco. Véase, por ejemplo, la discusión mantenida en el blog «Almacén de Derecho» entre los Profesores Alejandro Huergo Lora y Juli Ponce Solé en mayo de 2024, al calor de la Sentencia de la Audiencia Nacional denegatoria del acceso al código fuente: «Por qué aciertan las sentencias sobre el ‘algoritmo’ del bono social eléctrico» (<https://almacendederecho.org/por-que-aciertan-las-sentencias-sobre-el-algoritmo-del-bono-social-electrico>), «Por qué se equivocan las sentencias sobre el algoritmo del bono social eléctrico» (<https://almacendederecho.org/por-que-se-equivocan-las-sentencias-sobre-el-algoritmo-del-bono-social-electrico>), «Distopía y algoritmos» (<https://almacendederecho.org/distopia-y-algoritmos>) y «Utopía y algoritmos» (<https://almacendederecho.org/utopia-y-algoritmos>). La jurisprudencia, incluida la Sentencia del Tribunal Supremo que ha

El caso Bosco expone, en definitiva, el tipo de interferencia tecnológico-social que el CIAIS asume como objeto de estudio y para el que aspira a generar metodologías mixtas y, en buena medida, experimentales. Las dudas sobre la propia calificación del software, en el contexto de los procedimientos públicos de toma de decisiones, revelan cierto estado de inmadurez colectiva sobre el significado social de las tecnologías y el alcance de sus impactos. Se asume, por ejemplo, el prejuicio de que la codificación de una disposición jurídica en software es un mero ejercicio de traducción sintáctica, y no un verdadero salto entre distintos campos lógicos-lingüísticos, epistémicos, en realidad inconmensurables.

Decía Langdon Winner que nuestros aparatajes tecnológicos son auténticas «instituciones en desarrollo»¹⁸⁸. Aquellos significados e impactos han de observarse desde todas las aristas que exhibe esa institucionalidad tecnológicamente erigida. Perspectiva que ha de hacernos desconfiar del argumento que sostiene la presunta neutralidad de los objetos técnicos, que nos los presentas como meras herramientas «a la mano» (en expresión heideggeriana) susceptibles de «buenos» o «malos» usos que decida el sujeto. Como institutos sociales, las tecnologías alteran siempre los contextos en los que se insertan y recomponen al sujeto mismo (incluidos, por ejemplo, los operadores jurídicos) que, más que «usarlas», se acoplan o se ajustan a ellas.

Jurídicamente, una aplicación como Bosco reclama procedimientos de elaboración e implantación que no circunvalen los principios de publicidad y de participación pública asumidos desde largo como criterios irrenunciables en el proceder del poder público. Tecnológicamente, es necesario incorporar técnicas de desarrollo del *software* que habiliten aquellos principios jurídicos cumpliendo exigencias de trazabilidad e inteligibilidad. En nuestra coyuntura histórica, el Derecho tecnologizado de facto a de hacerlo también formalmente: urge encontrar acomodo al *software* decisorio en el sistema de fuentes del Derecho. Pero, al tiempo, la tecnología debe juridificarse, considerando que las exigencias que enuncia el Derecho son, también, criterios técnicos que condicionan su desarrollo, incluso al coste de modular ciertas utilidades, o excluirlas.

concedido finalmente el acceso al código fuente, han reiterado el carácter meramente auxiliar de Bosco. En una perspectiva, a nuestro juicio, más atinada, la Agencia Española de Protección de Datos que Bosco sí debe ser entendido como un sistema de decisión jurídica, al menos desde el punto de vista del tratamiento de datos de carácter personal que opera. De hecho, ha dictado resolución sancionadora contra la Administración Estatal por la inobservancia de los requisitos del art. 22 del Reglamento General de Protección de Datos [Reglamento (UE) 2016/679 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] con relación a los tratamientos automatizados de datos, incluida la elaboración de perfiles (Resolución de la AEPD de 6 de noviembre de 2024, PS-00324-2023, expediente sancionador Nº. 202211982).

¹⁸⁸ Winner, L. *La ballena y el reactor. Una búsqueda de los límites en la era de la alta tecnología*, Barcelona: Gedisa, 2008 (1ª ed. en inglés, 1986), pp. 55 y 101.

Ese quicio ilustra la propuesta metodológica del CIAIS. Esas hibridaciones disciplinares nos parecen irrenunciables en la pretensión de amarrar los desarrollos tecnológicos a exigencias sustantivas fruto de la deliberación colectiva. Definen, pues, un programa de preservación (y de reconfiguración inevitable) de los derechos humanos ante las inercias autorreferentes de la instrumentalidad tecnológica.

5. Conclusiones

1. El Colectivo de Inteligencia Artificial e Inteligencia Social (CIAIS) está compuesto por profesorado de la Universidad de La Rioja de muy diversas disciplinas. Sus miembros comparten una tesis teórica rotunda y elemental: la tecnología, en general, y la proliferación de los sistemas de inteligencia artificial, en particular, no tienen, en esencia, un sentido instrumental. De modo aplicado, sin duda, emergen tareas instrumentales episódicas, pero la tecnología -que despliega su existencia en una sociedad determinada- actúa en lo profundo como un transformador de las relaciones entre sujetos y objetos -y entre sujetos entre sí- modulando la manera de reconocer el mundo. Aun proyectada en una función instrumental, por tanto, la inteligencia artificial, esencialmente, preforma la realidad por el aparataje técnico en que revela el mundo. Los usos de la IA, más o menos bienintencionados, quedan condicionados por prescripciones inherentes al objeto técnico.
2. Sobre esta tesis de fondo se explica un desglose elemental de la tecnología informática apelando a la relación comunicativa entre el humano y la máquina: las proposiciones en lenguaje natural, en lenguaje jurídico o, en general, en lenguaje científico, generados en un contexto extralingüístico que apuntala su sentido, deben trasladarse a referencias formales que, en lenguaje informático, implican pulsos eléctricos, como puntos de difícil modulación contextual y solo codificables en lenguaje binario. La traducción, por tanto, del lenguaje natural al lenguaje formal informático (y viceversa) no es posible, ya que un mismo mensaje puede transcribirse en varias versiones de modo formalmente correcto, aunque sustantivamente distintos. La diferencia de granularidad de los lenguajes los hace inconmensurables.
3. La evolución tecnológica, así, siempre incorpora márgenes de incertidumbre, tanto en la creación de software como en su uso. Y esos márgenes, independientemente de las funciones instrumentales que el humano pretenda sobre la máquina, son una zona sin monitorizar. El despliegue de la tecnología desborda los criterios éticos, jurídicos o sociales. El riesgo, entonces, es que la supuesta neutralidad de la tecnología haga que reconozcamos un rumbo donde solo hay deriva. Sin embargo, reconocer esta realidad condiciona las posibilidades de interacción con la tecnología y con la IA, porque «somos capaces de experimentar lo técnico en su limitación».
4. En su aplicación como herramientas para la investigación o aplicación científica, los sistemas de IA se están mostrando imprescindibles en la carrera del imperativo tecnológico. Por competencia o colaboración entre grupos de investigación, por plazos esperables en la prestación de servicios y, en general, por armonización de usos sociales, cualquier experiencia científica necesita los atajos de la IA. Pero estos atajos tienen un costo en las claves expuestas. El CIAIS es un colectivo que aspira a

generar metodologías mixtas para desvelar los modos en que la inteligencia artificial, en distintos campos, deja rastro de pérdida de racionalidad social. Y si esta pérdida se observara solo desde la epistemología de la disciplina implicada, la ausencia de referencias impediría a los sujetos implicados reconocer el origen de la desorientación: ¿un mal uso de la tecnología?

5. La generación de metodologías mixtas desarrolladas en el CIAIS, puestas en práctica a través de un amplio elenco de actividades y ponencias, pretende que el funcionamiento de la IA en un determinado ámbito o disciplina (la salud, el trabajo, la docencia, la política, la comunicación, el medioambiente, etc.) quede expuesto a una evaluación de racionalidad cruzada en la que poder detectar la pérdida de justificación social. Las bases teóricas que ponen de relieve los riesgos de la tecnología, por tanto, se activan en escenarios concretos en los que reconocer y evitar o paliar el desvalor social de la mediación por sistemas de IA.
6. Que estas mermas de racionalidad impliquen, incluso, afectación a derechos humanos hace de la mirada jurídica una epistemología transversal a cualquier ámbito. La alerta desde el Derecho ha de ser, en las claves expuestas, una exhortación a la prudencia en el uso de sistemas de IA.

Bibliografía

Agencia de los Derechos Fundamentales de la Unión Europea (FRA). *Trabajadores de plataformas digitales y derechos fundamentales*. Luxemburgo: Publications Office of the European Union, 2021.

Ayer, A. J. (comp.). *El positivismo lógico*. México: Fondo de Cultura Económica, 1981.

Chomsky, N. *Aspectos de la teoría de la sintaxis*. Madrid: Aguilar, 1970.

Consejo de Europa. *Recommendation CM/Rec(2020)1 on the human rights impacts of algorithmic systems*. Estrasburgo: Council of Europe, 2020.

Floridi, L. *Filosofia dell'informazione*. Milano: Raffaello Cortina Editore, 2024.

Floridi, L. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford: Oxford University Press, 2014.

García de Enterría, E. *La lengua de los derechos. La formación del Derecho Público europeo tras la Revolución Francesa*, Madrid: Alianza, 1994.

Heidegger, M. «La pregunta por la técnica». En: *Conferencias y artículos*. Barcelona: Ediciones del Serbal, 1994.

Hui, Y. *La cuestión de la tecnología en China: Un ensayo sobre cosmotécnica*. Buenos Aires: Caja Negra Editora, 2016.

Kuhn, T. S. *La estructura de las revoluciones científicas*. 4ª ed. México: Fondo de Cultura Económica (Breviarios), 2017.

Ortega y Gasset, J. «Meditación de la técnica». En: *Obras completas*. Vol. V. Madrid: Revista de Occidente, 1964, pp. 317-375.

Searle, J. R. *Mentes, cerebros y ciencia*. Madrid: Editorial Cátedra, 2001.

Simondon, G. *El modo de existencia de los objetos técnicos*. Buenos Aires: Prometeo Libros, 2013.

Tarde, G. *Les lois de l'imitation. Étude sociologique*. Paris: Félix Alcan éditeur, 1890.

Thiel, C. *Sentido y referencia en la lógica de Gottlob Frege*. Madrid: Editorial Tecnos, 1972.

Weber, M. *Economía y Sociedad*, México D.F.: Fondo de Cultura Económica, 2014.

Winner, L. *La ballena y el reactor. Una búsqueda de los límites en la era de la alta tecnología*, Barcelona: Gedisa, 2008.

Wittgenstein, L. *Tractatus logico-philosophicus*. Madrid: Tecnos, 2017.

Wittgenstein, L. *Investigaciones filosóficas*. Madrid: Editorial Trotta, 2021.

Zuboff, S. *The Age of Surveillance Capitalism*. New York: PublicAffairs, 2019.

Paula Sáez Álvarez. Voluntary Carbon Credits y activos digitales hacia una gobernanza global basada en derechos¹⁸⁹

Paula Sáez Álvarez

Profesora Permanente Laboral

Universidad CEU Fernando III, CEU Universities

[ACCEDE A LA GRABACIÓN](#)

Resumen

El presente trabajo analiza la naturaleza jurídica de los créditos de carbono voluntarios (*Verified Carbon Credits*, VCCs) desde la perspectiva del Derecho privado. Estos instrumentos, concebidos para certificar reducciones verificadas de emisiones de gases de efecto invernadero, han proliferado bajo esquemas privados y fragmentados, al margen de los mercados regulados creados en el Protocolo de Kioto y el Acuerdo de París. Si bien los VCCs han contribuido a canalizar financiación hacia proyectos de mitigación, su falta de uniformidad normativa ha generado riesgos de doble contabilidad, altos costes de transacción e inseguridad jurídica.

Este trabajo examina tanto el ámbito jurídico de estos instrumentos, aún por explorar, como la interacción entre VCCs, justicia climática y desarrollo del Sur Global. Aunque los proyectos voluntarios pueden generar beneficios directos en comunidades locales —acceso a energía, reforestación o reducción de costes— también plantean riesgos de desigualdad, apropiación por élites o empleo precario. De ahí la necesidad de acompañar la regulación con salvaguardas sociales y ambientales efectivas. Finalmente, se consideran los aportes de la tecnología *blockchain* y la inteligencia artificial para reforzar la trazabilidad y la transparencia, siempre que se integren en infraestructuras jurídicas robustas.

¹⁸⁹ Esta publicación se realiza en el marco del Proyecto «Eficiencia tecnológica, dignidad y derechos humanos. ¿Quo vadis, humanidad, en la era de la IA?», Convocatoria de Ayudas a Proyectos de Investigación en Líneas Estratégicas CEU 2024-2025, Coordinación de Universidades CEU (Fundación Universitaria San Pablo CEU). La realización de este trabajo trae como causa la estancia de investigación realizada en el Instituto Internacional para la Unificación del Derecho Privado (UNIDROIT) durante el año 2023.

El trabajo concluye que los VCCs solo alcanzarán legitimidad si cumplen simultáneamente tres exigencias: integridad climática, seguridad jurídica y justicia social. La armonización normativa propuesta por UNIDROIT representa un paso esencial hacia un mercado de créditos voluntarios confiable, capaz de alinear eficiencia económica, confianza jurídica y equidad en la transición energética global.

Introducción

Desde comienzos del siglo XXI se observa un incremento sostenido de las emisiones de gases de efecto invernadero, con un peso creciente del CO₂ procedente de economías emergentes. Pese a los acuerdos internacionales, las emisiones globales continúan aumentando, lo que ha impulsado la creación de instrumentos de mercado —tanto regulados como voluntarios— destinados a canalizar financiación hacia proyectos de mitigación y acelerar la reducción efectiva de emisiones. El cambio climático constituye uno de los mayores desafíos del siglo XXI, al poner en riesgo la sostenibilidad ambiental, económica y social a nivel global, no siendo, el Derecho mercantil ajeno a estos retos.

En este contexto, los créditos de carbono se han consolidado como instrumentos clave para obtener financiación en proyectos de mitigación y para facilitar el cumplimiento de compromisos de reducción de emisiones. Junto a los mercados regulados —vinculados al Protocolo de Kioto y al Acuerdo de París— han proliferado los mercados voluntarios de carbono, en los que la voluntad de las partes se pone de manifiesto al adquirir créditos verificados con fines de compensación y de responsabilidad social corporativa.

Los créditos de carbono voluntarios (*Verified Carbon Credits*, VCCs) – también denominados bonos de carbono – representan una certificación que acredita que el titular, directa o indirectamente, ha reducido o eliminado de la atmósfera una tonelada métrica de dióxido de carbono¹⁹⁰, de conformidad con las normas y requisitos aplicables según lo verificado por un tercero. Los VCCs representan reducciones de emisiones certificadas por entidades privadas, cuya negociación se ha multiplicado y realizado de diversas formas en ausencia de un marco jurídico internacional uniforme. Este vacío genera problemas de seguridad jurídica, riesgo de doble contabilidad y diversidad de estándares, lo que afecta tanto a la implantación internacional del instrumento como a su eficacia (jurídica y climática). De ahí la importancia de examinar su naturaleza jurídica y su tratamiento como activo digital intangible, susceptible de ser transmitido, gravado o utilizados como garantía en el tráfico jurídico.

El interés de esta investigación reside en la necesidad de aportar claridad conceptual y soluciones normativas que fortalezcan la gobernanza global de los mercados voluntarios. El análisis se centra en el Proyecto de UNIDROIT sobre VCCs, que busca dotar de principios uniformes de Derecho privado a estas transacciones, en diálogo con los *Principles on Digital Assets and Private Law* de 2023. El trabajo aborda, asimismo, la interacción entre VCCs, nuevas

¹⁹⁰ Un bono de carbono equivale a una tonelada de CO₂ (2).

tecnologías —como *blockchain* o inteligencia artificial— y la dimensión de justicia climática en el Sur Global.

El trabajo se estructura en varias partes. En primer lugar, se revisa la evolución de los mercados de carbono y el surgimiento de los VCCs como instrumentos diferenciados. En segundo lugar, se examina su consideración como bienes mercantiles y activos digitales, con especial referencia al análisis de UNIDROIT en su trabajo. En tercer lugar, se analizan experiencias nacionales y regionales, en especial en la Unión Europea, con especial interés en Portugal. Finalmente, se aborda la conexión entre VCCs, derechos humanos y justicia climática, con atención al impacto en comunidades del Sur Global.

La finalidad última de este estudio es contribuir a la construcción de un marco jurídico coherente y previsible que permita a los VCCs cumplir su doble función: reforzar la integridad climática y generar confianza en los mercados internacionales, alineando eficiencia económica, seguridad jurídica y justicia social.

Los créditos de carbono: entre compensación y activo digital

Evolución del mercado de carbono.

El mercado internacional de carbono comenzó a consolidarse con instrumentos como el Protocolo de Kioto (1997) y el Acuerdo de París (2015), que fomentaron la creación y el comercio de certificados de reducción de emisiones.

El Protocolo de Kioto de 1997, adoptado en el marco de la Convención Marco de Naciones Unidas sobre el Cambio Climático (CMNUCC)¹⁹¹, complementó la Convención de 1992¹⁹², introduciendo por primera vez mecanismos de que permitían a los Estados industrializados dar cumplimiento a sus compromisos de reducción de emisiones mediante fórmulas basadas en el mercado. El Protocolo de Kyoto se erigió sobre principios clásicos del Derecho ambiental: acción preventiva, precaución y responsabilidades comunes pero diferenciadas. Estos principios legitiman la creación de incentivos económicos para los países industrializados y, al mismo tiempo, de mecanismos de cooperación que permiten a los países en desarrollo participar en proyectos de reducción con compensación económica ⁽¹⁾.

¹⁹¹ Texto disponible en [convsp.pdf](#).

¹⁹² Junto con junto a la [Convención de Naciones Unidas sobre Diversidad Biológica](#) y la [Convención de Naciones Unidas de Lucha contra la Desertificación](#).

Entre los principales mecanismos de Kioto destacan tres. En primer lugar, el artículo 6 reguló la denominada Aplicación Conjunta (*Joint Implementation*)¹⁹³, que habilitaba a las Partes incluidas en el Anexo I del Protocolo a transferir o adquirir unidades de reducción de emisiones (UREs) resultantes de proyectos encaminados a reducir emisiones o incrementar absorciones en sectores de la economía de otras Partes del mismo grupo. Estos proyectos debían cumplir requisitos específicos, como la adicionalidad de las reducciones (es decir, que fueran superiores a las que se producirían en ausencia del proyecto) y la aprobación previa por las Partes participantes. Asimismo, el artículo exigía que la adquisición de unidades fuese suplementaria a las medidas nacionales, con el fin de evitar que el comercio sustituyera por completo cualesquiera otras políticas internas de mitigación del cambio climático.

En segundo lugar, el artículo 12 introdujo el Mecanismo de Desarrollo Limpio (MDL), concebido para articular la cooperación entre países industrializados y países en desarrollo¹⁹⁴. Bajo este

193 Artículo 6 del Protocolo de Kioto “1. A los efectos de cumplir los compromisos contraídos en virtud del artículo 3, toda Parte incluida en el anexo I podrá transferir a cualquiera otra de esas Partes, o adquirir de ella, las unidades de reducción de emisiones resultantes de proyectos encaminados a reducir las emisiones antropógenas por las fuentes o incrementar la absorción antropógena por los sumideros de los gases de efecto invernadero en cualquier sector de la economía, con sujeción a lo siguiente: a) Todo proyecto de ese tipo deberá ser aprobado por las Partes participantes; b) Todo proyecto de ese tipo permitirá una reducción de las emisiones por las fuentes, o un incremento de la absorción por los sumideros, que sea adicional a cualquier otra reducción u otro incremento que se produciría de no realizarse el proyecto; c) La Parte interesada no podrá adquirir ninguna unidad de reducción de emisiones si no ha dado cumplimiento a sus obligaciones dimanantes de los artículos 5 y 7; y d) La adquisición de unidades de reducción de emisiones será suplementaria a las medidas nacionales adoptadas a los efectos de cumplir los compromisos contraídos en virtud del artículo 3. 2. La Conferencia de las Partes en calidad de reunión de las Partes en el presente Protocolo podrá, en su primer período de sesiones o tan pronto como sea posible después de éste, establecer otras directrices para la aplicación del presente artículo, en particular a los efectos de la verificación y presentación de informes. 3. Una Parte incluida en el anexo I podrá autorizar a personas jurídicas a que participen, bajo la responsabilidad de esa Parte, en acciones conducentes a la generación, transferencia o adquisición en virtud de este artículo de unidades de reducción de emisiones. Si, de conformidad con las disposiciones pertinentes del artículo 8, se plantea alguna cuestión sobre el cumplimiento por una Parte incluida en el anexo I de las exigencias a que se refiere el presente artículo, la transferencia y adquisición de unidades de reducción de emisiones podrán continuar después de planteada esa cuestión, pero ninguna Parte podrá utilizar esas unidades a los efectos de cumplir sus compromisos contraídos en virtud del artículo 3 mientras no se resuelva la cuestión del cumplimiento.”

194 Artículo 12 del Protocolo de Kioto “1. Por el presente se define un mecanismo para un desarrollo limpio. 2. El propósito del mecanismo para un desarrollo limpio es ayudar a las Partes no incluidas en el anexo I a lograr un desarrollo sostenible y contribuir al objetivo último de la Convención, así como ayudar a las Partes incluidas en el anexo I a dar cumplimiento a sus compromisos cuantificados de limitación y reducción de las emisiones contraídos en virtud del artículo 3. 3. En el marco del mecanismo para un desarrollo limpio: a) Las Partes no incluidas en el anexo I se beneficiarán de las actividades de proyectos que tengan por resultado reducciones certificadas de las emisiones; y b) Las Partes incluidas en el anexo I podrán utilizar las reducciones certificadas de emisiones resultantes de esas actividades de proyectos para contribuir al cumplimiento de una parte de sus compromisos cuantificados de limitación y reducción de las emisiones contraídos en virtud del artículo 3, conforme lo determine la Conferencia de las Partes en calidad de reunión de las Partes en el presente Protocolo. 4. El mecanismo para un desarrollo limpio estará sujeto a la autoridad y la dirección de la Conferencia de las Partes en calidad de reunión de las Partes en el presente Protocolo y a la supervisión de una junta ejecutiva del mecanismo para un desarrollo limpio. 5. La reducción de emisiones resultante de cada actividad de proyecto deberá ser certificada por las entidades operacionales que designe la Conferencia de las Partes en calidad de reunión de las Partes en el presente Protocolo sobre la base de: a) La participación voluntaria acordada por cada Parte participante; b) Unos beneficios reales, mensurables y a largo plazo en relación con la mitigación del cambio climático; y c) Reducciones de las emisiones que sean adicionales a las que se producirían en ausencia de la actividad de proyecto certificada. 6. El mecanismo para un desarrollo limpio ayudará según sea necesario a organizar la

esquema, las Partes no incluidas en el Anexo I podían beneficiarse de proyectos que contribuyeran al desarrollo sostenible (a través de capturar el CO₂ de la atmósfera, mitigar las emisiones de gases efecto invernadero o contrarrestar los efectos del calentamiento global (2)) y generaran Reducciones Certificadas de Emisiones (RCE), mientras que las Partes del Anexo I podían utilizar dichas reducciones para cumplir parte de sus compromisos. Por ello, este instrumento se ha considerado un mecanismo atractivo tanto para los países industrializados como para los países en desarrollo (3). El MDL estuvo sometido a la autoridad de la Conferencia de las Partes y a la supervisión de una Junta Ejecutiva, con exigencias de verificación independiente y criterios de transparencia, eficiencia y rendición de cuentas. En el marco de esta herramienta, los países industrializados podían adquirir las reducciones certificadas de emisiones generadas por proyectos en países en desarrollo y contabilizarlas como parte de sus propios compromisos de reducción. Esto ocurría sin que el país anfitrión tuviera que ajustar sus cifras nacionales, es decir, las reducciones se sumaban al cumplimiento del comprador sin restarse de los objetivos del país donde se realizaba el proyecto. De este modo, la acción de mitigación beneficiaba principalmente al país que adquiría las reducciones, mientras que el país anfitrión recibía incentivos económicos por su participación, pero no veía modificada su contabilidad oficial de emisiones.

El artículo 17 añadió un tercer mecanismo¹⁹⁵: el comercio internacional de emisiones, mediante el cual las Partes del Anexo B podían participar en operaciones de compraventa de derechos de emisión como complemento a sus políticas nacionales¹⁹⁶. Este último supuso la auténtica

financiación de actividades de proyectos certificadas. 7. La Conferencia de las Partes en calidad de reunión de las Partes en el presente Protocolo en su primer período de sesiones deberá establecer las modalidades y procedimientos que permitan asegurar la transparencia, la eficiencia y la rendición de cuentas por medio de una auditoría y la verificación independiente de las actividades de proyectos. 8. La Conferencia de las Partes en calidad de reunión de las Partes en el presente Protocolo se asegurará de que una parte de los fondos procedentes de las actividades de proyectos certificadas se utilice para cubrir los gastos administrativos y ayudar a las Partes que son países en desarrollo particularmente vulnerables a los efectos adversos del cambio climático a hacer frente a los costos de la adaptación. 9. Podrán participar en el mecanismo para un desarrollo limpio, en particular en las actividades mencionadas en el inciso a) del párrafo 3 supra y en la adquisición de unidades certificadas de reducción de emisiones, entidades privadas o públicas, y esa participación quedará sujeta a las directrices que imparta la junta ejecutiva del mecanismo para un desarrollo limpio. 10. Las reducciones certificadas de emisiones que se obtengan en el período comprendido entre el año 2000 y el comienzo del primer período de compromiso podrán utilizarse para contribuir al cumplimiento en el primer período de compromiso.”

¹⁹⁵ Artículo 17 del Protocolo de Kioto “La Conferencia de las Partes determinará los principios, modalidades, normas y directrices pertinentes, en particular para la verificación, la presentación de informes y la rendición de cuentas en relación con el comercio de los derechos de emisión. Las Partes incluidas en el anexo B podrán participar en operaciones de comercio de los derechos de emisión a los efectos de cumplir sus compromisos dimanantes del artículo 3. Toda operación de este tipo será suplementaria a las medidas nacionales que se adopten para cumplir los compromisos cuantificados de limitación y reducción de las emisiones dimanantes de ese artículo.”

¹⁹⁶ De forma complementaria a los anteriores, los apartados 10 y 11 del artículo 3 establecieron la arquitectura contable de estos mecanismos, garantizando su integridad ambiental. Toda unidad de reducción de emisiones adquirida conforme a los artículos 6 o 17 debía sumarse a la cantidad atribuida al país comprador, mientras que toda unidad transferida debía deducirse de la cantidad atribuida al país vendedor. Este diseño aseguraba un sistema de “suma cero”, evitando la creación artificial de reducciones y reforzando la credibilidad del régimen.

“sorpresa de Kyoto”, al abrir un cauce de cooperación Norte-Sur que se convirtió en la modalidad más relevante en términos prácticos (1).

En conjunto, estos preceptos del Protocolo de Kioto configuraron un mercado internacional de carbono regulado, en el que coexistían proyectos bilaterales entre países industrializados (AC), proyectos de cooperación con países en desarrollo (MDL) y el comercio de derechos de emisión entre Estados. Su puesta en marcha no solo abrió la puerta a una nueva forma de cooperación climática, sino que también sentó las bases conceptuales para los posteriores mercados voluntarios de carbono, ya que admitió que el mercado podía ayudar en el proceso de emisión de gases de efecto invernadero mediante la propuesta de crear un valor negociable para estas reducciones (3).

La experiencia de Kioto evidenció ciertas debilidades de los mecanismos de mercado, como la existencia de créditos de “*hot air*” procedentes de la antigua Unión Soviética, que minaron la integridad ambiental del sistema, y la complejidad institucional del MDL, que elevó los costes de transacción. Los créditos conocidos como *hot air* o “aire caliente” son unidades de emisiones que un país del Anexo I puede vender cuando sus emisiones se encuentran por debajo de las metas fijadas, incluso sin haber realizado esfuerzos adicionales para reducirlas (1). Esto ocurre, por ejemplo, en países cuyas metas son relativamente bajas o cuyos niveles de emisiones habían disminuido por razones económicas, como sucedió en la antigua Unión Soviética tras su recesión, lo que generó un gran excedente de emisiones acreditables. A diferencia de los créditos generados por proyectos de Mecanismo de Desarrollo Limpio (MDL), que representan reducciones reales y verificables en países en desarrollo, el Hot Air no se asocia a proyectos concretos de mitigación, sino que se aplica únicamente en el contexto de Implementación Conjunta o Comercio de Emisiones entre países industrializados. Por ello, aunque estos créditos pueden utilizarse para cumplir metas climáticas, no implican necesariamente acciones adicionales que reduzcan realmente los gases de efecto invernadero, lo que les otorga su carácter de “aire caliente”.

Al mismo tiempo, la complejidad institucional del MDL, con múltiples actores (Junta Ejecutiva, entidades operacionales, autoridades nacionales designadas), elevó los costes de transacción. Estas limitaciones explican por qué el tránsito al Acuerdo de París buscó simplificar la arquitectura y ampliar la participación de todos los países bajo las NDCs, esto es, Contribuciones Determinadas a Nivel Nacional (1). El tránsito de Kioto a París transformó la arquitectura de cooperación. El sistema pasó de un MDL focalizado en países del Anexo I a un esquema universal de NDCs, que exigió evitar la doble contabilidad y alinear mercados y registros con trayectorias compatibles con la meta mantener el aumento de la temperatura media global muy por debajo de 2 °C respecto a niveles preindustriales, e intentar limitarlo a 1,5 °C. Este giro amplió la base de participantes y exigió una coordinación técnica y jurídica más fina entre registros nacionales y estándares privados (4).

Con París, la lógica cambió: todas las Partes presentaron Contribuciones Determinadas a nivel Nacional (NDCs) y cualquier transferencia internacional requirió ajustes para evitar doble contabilidad. A diferencia del Protocolo de Kioto, que imponía obligaciones diferenciadas según la pertenencia al Anexo I, el Acuerdo de París se articuló sobre las NDCs, de alcance universal, en las que cada Estado asumía compromisos de reducción de emisiones conforme a sus capacidades y circunstancias nacionales.

El artículo 6 del Acuerdo de París constituye el núcleo de la regulación de los mecanismos de cooperación y mercados de carbono en este nuevo marco¹⁹⁷. En primer lugar, reconoce que las Partes pueden cooperar voluntariamente en la aplicación de sus NDC, a fin de incrementar la ambición en materia de mitigación y adaptación, promoviendo al mismo tiempo el desarrollo sostenible y la integridad ambiental, ex artículo 6.1 (5). Para ello, el acuerdo contempla la posibilidad de recurrir a enfoques cooperativos que impliquen la transferencia internacional de resultados de mitigación (*Internationally Transferred Mitigation Outcomes*, ITMO). En segundo lugar, el Acuerdo crea un mecanismo específico de desarrollo sostenible (artículo 6, apartados 4-7), sucesor del MDL del Protocolo de Kioto, con un mayor alcance. Sus objetivos son promover la mitigación de emisiones, facilitar la participación de entidades públicas y privadas,

¹⁹⁷ Artículo 6 del Acuerdo de París “Las Partes reconocen que algunas Partes podrán optar por cooperar voluntariamente en la aplicación de sus contribuciones determinadas a nivel nacional para lograr una mayor ambición en sus medidas de mitigación y adaptación y promover el desarrollo sostenible y la integridad ambiental. 2. Cuando participen voluntariamente en enfoques cooperativos que entrañen el uso de resultados de mitigación de transferencia internacional para cumplir con las contribuciones determinadas a nivel nacional, las Partes deberán promover el desarrollo sostenible y garantizar la integridad ambiental y la transparencia, también en la gobernanza, y aplicar una contabilidad robusta que asegure, entre otras cosas, la ausencia de doble cómputo, de conformidad con las orientaciones que haya impartido la Conferencia de las Partes en calidad de reunión de las Partes en el presente Acuerdo. 3. La utilización de resultados de mitigación de transferencia internacional para cumplir con las contribuciones determinadas a nivel nacional en virtud del presente Acuerdo será voluntaria y deberá ser autorizada por las Partes participantes. 4. Por el presente se establece un mecanismo para contribuir a la mitigación de las emisiones de gases de efecto invernadero y apoyar el desarrollo sostenible, que funcionará bajo la autoridad y la orientación de la Conferencia de las Partes en calidad de reunión de las Partes en el presente Acuerdo y podrá ser utilizado por las Partes a título voluntario. El mecanismo será supervisado por un órgano que designará la Conferencia de las Partes en calidad de reunión de las Partes en el presente Acuerdo, y tendrá por objeto: a) Promover la mitigación de las emisiones de gases de efecto invernadero, fomentando al mismo tiempo el desarrollo sostenible; b) Incentivar y facilitar la participación, en la mitigación de las emisiones de gases de efecto invernadero, de las entidades públicas y privadas que cuenten con la autorización de las Partes; c) Contribuir a la reducción de los niveles de emisión en las Partes de acogida, que se beneficiarán de actividades de mitigación por las que se generarán reducciones de las emisiones que podrá utilizar también otra Parte para cumplir con su contribución determinada a nivel nacional; y d) 5. Producir una mitigación global de las emisiones mundiales. Las reducciones de las emisiones que genere el mecanismo a que se refiere el párrafo 4 del presente artículo no deberán utilizarse para demostrar el cumplimiento de la contribución determinada a nivel nacional de la Parte de acogida, si otra Parte las utiliza para demostrar el cumplimiento de su propia contribución determinada a nivel nacional. 6. La Conferencia de las Partes en calidad de reunión de las Partes en el presente Acuerdo velará por que una parte de los fondos devengados de las actividades que se realicen en el marco del mecanismo a que se refiere el párrafo 4 del presente artículo se utilice para sufragar los gastos administrativos y para ayudar a las Partes que son países en desarrollo particularmente vulnerables a los efectos adversos del cambio climático a hacer frente a los costos de la adaptación. 7. La Conferencia de las Partes en calidad de reunión de las Partes en el presente Acuerdo aprobará las normas, las modalidades y los procedimientos del mecanismo a que se refiere el párrafo 4 del presente artículo en su primer período de sesiones. 8. Las Partes reconocen la importancia de disponer de enfoques no relacionados con el mercado que sean integrados, holísticos y equilibrados y que les ayuden a implementar sus contribuciones determinadas a nivel nacional, en el contexto del desarrollo sostenible y de la erradicación de la pobreza y de manera coordinada y eficaz, entre otras cosas mediante la mitigación, la adaptación, la financiación, la transferencia de tecnología y el fomento de la capacidad, según proceda. Estos enfoques tendrán por objeto: a) b) Promover la ambición relativa a la mitigación y la adaptación; Aumentar la participación de los sectores público y privado en la aplicación de las contribuciones determinadas a nivel nacional; y c) Ofrecer oportunidades para la coordinación de los instrumentos y los arreglos institucionales pertinentes. 9. Por el presente se define un marco para los enfoques de desarrollo sostenible no relacionados con el mercado, a fin de promover los enfoques no relacionados con el mercado a que se refiere el párrafo 8 del presente artículo.”

beneficiar a las Partes de acogida mediante reducciones certificadas y, sobre todo, asegurar una mitigación global neta de emisiones. Además, el acuerdo prevé que parte de los ingresos generados se destinen a apoyar a los países en desarrollo más vulnerables a los efectos del cambio climático, reforzando así la dimensión de justicia climática¹⁹⁸.

Mientras los mercados regulados de carbono (como el MDL o el EU ETS, objeto de estudio más adelante) responden a obligaciones internacionales vinculantes, los mercados voluntarios surgieron al margen de tales compromisos, con el objetivo de permitir a empresas y particulares compensar emisiones sin estar sujetos a cuotas legales. La literatura destaca que estos esquemas voluntarios se han multiplicado de manera descentralizada, bajo un mosaico de estándares privados no coordinados, lo que explica la diversidad de reglas de certificación y verificación presentes en la actualidad (6). Los mercados voluntarios de carbono, a diferencia de los de cumplimiento, no se fundamentan en obligaciones regulatorias, sino en acciones voluntarias de empresas, ONG o individuos (7). En muchos casos, la motivación principal es la responsabilidad social corporativa o la anticipación de futuras regulaciones. Esto aumenta la fragmentación y genera cierta inseguridad jurídica, aunque permite a empresas y particulares compensar emisiones fuera de las obligaciones legales.

Los créditos de carbono como bienes mercantiles

Los mecanismos de flexibilidad introducidos por el Protocolo de Kioto y posteriormente desarrollados en el Acuerdo de París dieron lugar a una serie de instrumentos jurídicos estandarizados, concebidos para representar reducciones verificadas de emisiones de gases de efecto invernadero. Con ello se configuró un conjunto de unidades estandarizadas (AAU, CER, EUA, entre otras) de reducción de emisiones con una base normativa y procedimientos de verificación internacionalmente reconocidos.

En el marco de Kioto, los principales instrumentos fueron los siguientes. En primer lugar, hallamos los *Certified Emission Reductions* (CERs), conocidos como “bonos de carbono”, que constituyen unidades derivadas del MDL (artículo 12 del Protocolo), generadas por proyectos de reducción o absorción en países no incluidos en el Anexo I. Su emisión estaba sujeta a la supervisión de la Junta Ejecutiva del MDL y a procedimientos estrictos de validación y verificación. Los certificados de reducción de emisiones (CERs) constituyen derechos transferibles que representan una tonelada de CO₂ equivalente reducida o capturada. Su

¹⁹⁸ Además, el artículo 6 introduce un marco para enfoques no basados en el mercado (artículo 6, apartados 8–9). Estos enfoques buscan ofrecer alternativas integradas y holísticas que contribuyan a la mitigación, la adaptación, la transferencia tecnológica y el fortalecimiento de capacidades, sin necesidad de generar créditos comercializables. Se pretende así complementar los instrumentos de mercado con mecanismos de cooperación más amplios, orientados al desarrollo sostenible y la erradicación de la pobreza.

negociación en mercados organizados y su transferencia mediante contratos de compraventa (ERPAs) anticiparon la lógica mercantil de los actuales créditos voluntarios: bienes intangibles con valor económico, cuya eficacia depende de su adicionalidad, verificabilidad y registro en una infraestructura reconocida (1).

En segundo lugar, las Unidades de Reducción de Emisiones (ERUs, *Emission Reduction Units*), surgen de la Aplicación Conjunta del artículo 6 del Protocolo, que representan las reducciones logradas en proyectos realizados en países incluidos en el Anexo I. A diferencia de los CERs, estas unidades se intercambiaban únicamente entre Estados con compromisos cuantificados de reducción.

En tercer lugar, el Protocolo de Kioto introdujo asimismo las *Assigned Amount Units* (AAUs), ligadas al artículo 3 del Protocolo de Kioto, que representaban la cantidad total de emisiones permitidas a cada Estado en virtud de sus compromisos. Estas unidades eran transferibles en el marco del comercio internacional de emisiones (el ya mencionado artículo 17), funcionando en la práctica como “derechos de emisión” cuya negociación permitía ajustar el cumplimiento de los objetivos de cada país.

Más adelante, el Acuerdo de París introdujo una nueva categoría: los *Internationally Transferred Mitigation Outcomes* (ITMOs), previstos en su artículo 6. Estos buscaban evitar la doble contabilidad, funcionando como sucesores de los CER y ERU. A diferencia de las unidades del régimen de Kioto, los ITMOs podían originarse en cualquier Parte del Acuerdo, sin distinción entre países desarrollados y en desarrollo, siempre que cumpliesen con las reglas de transparencia, integridad ambiental y contabilidad.

En el marco de la Unión Europea, destacan las *EU-Allowances* (EUAs), propias del Sistema de Comercio de Emisiones de la Unión Europea (EU ETS). Funcionan como permisos de emisión emitidos bajo un sistema de tope y comercio (*cap-and-trade*)¹⁹⁹, lo que garantiza un control centralizado de la cantidad total de emisiones permitidas y otorga a estas unidades un valor de mercado dinámico.

Desde una óptica mercantil, tanto las unidades de Kioto (CERs, ERUs, AAUs) como los ITMO del Acuerdo de París operan bajo la lógica de bienes intangibles registrados electrónicamente —

¹⁹⁹ Un sistema de tope y comercio (*cap-and-trade*) es un mecanismo de mercado en el que la autoridad fija un límite máximo de emisiones (*cap*) y reparte permisos equivalentes a una cantidad determinada de CO₂ (19). Estos permisos pueden comprarse y venderse (*trade*) entre las empresas participantes, de modo que el límite global se mantiene mientras cada actor decide si reducir emisiones o adquirir derechos adicionales.

que los Principios de UNIDROIT califican como activos digitales—, con un valor de mercado determinado.

La lógica de activo transmisible es la que, posteriormente, servirá de referencia para el desarrollo de los créditos de carbono voluntarios (VCCs), cuya diferencia principal radica en que su emisión y validación no dependen de un régimen internacional vinculante, sino de esquemas privados de certificación. Los mercados voluntarios de carbono han proliferado bajo un mosaico de estándares privados no coordinados, lo que explica la diversidad de reglas de certificación y verificación presentes en la actualidad²⁰⁰. Entre los principales estándares se encuentran:

- *Verified Carbon Standard* (VCS): El programa de certificación de gases de efecto invernadero más utilizado a nivel mundial. Los proyectos bajo este estándar han reducido o eliminado más de mil millones de toneladas de emisiones de gases de efecto invernadero de la atmósfera.
- *Gold Standard*: Establecido por WWF y otras organizaciones, se centra en proyectos que contribuyen al desarrollo sostenible además de la reducción de emisiones.
- *Plan Vivo*: Orientado a proyectos comunitarios que promueven la gestión sostenible de los recursos naturales.
- *Climate Action Reserve* (CAR) y *American Carbon Registry* (ACR): Estándares utilizados principalmente en América del Norte.
- *Label Bas Carbone*: Estándar francés que promueve proyectos de captura de carbono en el sector agrícola y forestal.

Las primeras se corresponden con los *Voluntary Carbon Credits* (VCCs), como instrumentos creados y verificados al margen de la regulación internacional, pero con una lógica muy similar a la de los créditos oficiales. Se trata de certificados que representan reducciones de emisiones validadas por entidades privadas²⁰¹, cuya finalidad principal es responder a compromisos empresariales de responsabilidad social, marketing verde o estrategias de *net-zero*.

La distinción entre mercados de cumplimiento y mercados voluntarios permite comprender la evolución de las diferentes unidades de carbono y su carácter mercantil (2). En los primeros se ubican los ya mencionados instrumentos creados por el Protocolo de Kioto —AAU, CER y ERU—, junto con los sistemas regionales posteriores como el EU ETS y sus EUA, todos ellos sustentados en obligaciones jurídicas y en un marco regulatorio vinculante. En cambio, los mercados voluntarios se apoyan en esquemas no sujetos a obligaciones estatales, bien a través

²⁰⁰ En un nivel intermedio se sitúan las *Emission Reduction Units* (ERUs), procedentes de proyectos de Aplicación Conjunta regulados por el artículo 6 del Protocolo de Kioto. Estas unidades, surgidas de reducciones verificadas en países del Anexo I del Protocolo, se basan en estándares aprobados por los gobiernos participantes, lo que refuerza su fiabilidad y legitimidad en comparación con los créditos voluntarios.

²⁰¹ Como *Verified Carbon Standard*, el programa de créditos de gases de efecto invernadero más utilizado en el mundo.

de plataformas oficiales de negociación, bien bajo estándares privados y autorregulados. Estos últimos han dado lugar a instrumentos como los VCCs. La coexistencia de ambos ecosistemas refleja la progresiva mercantilización de las reducciones de emisiones: desde instrumentos internacionales regulados y con reconocimiento normativo hasta créditos generados bajo paraguas voluntarios y privados, cuyo origen responde más a motivaciones corporativas o reputacionales que a obligaciones jurídicas provenientes de pactos o convenciones internacionales.

Como hemos adelantado, un bono de carbono equivale a una tonelada de CO₂ reducida o absorbida (2), que puede ser negociado en el mercado internacional, lo que dio lugar a la creación de un mercado específico para la reducción de gases de efecto invernadero (GEI). El funcionamiento responde, básicamente, a una lógica de oferta y demanda regulada por cuotas de emisión. Los acuerdos internacionales, como el Protocolo de Kioto, fijaron un límite máximo de emisiones para los países industrializados, mientras que la normativa nacional concreta estos objetivos a través de cuotas asignadas a las industrias. En consecuencia, las empresas que no alcanzan sus objetivos de reducción han optado por adquirir en el mercado la cuota de contaminación que otra empresa —o incluso otro país— no ha emitido. Por el contrario, aquellas que han logrado reducir sus emisiones por debajo de los niveles exigidos pueden vender el excedente de reducción o el “permiso de emisión” sobrante en los mercados nacionales o internacionales. Este funcionamiento pone de manifiesto que los créditos de carbono operan como activos mercantiles, susceptibles de ser transmitidos, pignorados, cedidos, etc.

Un ejemplo ilustrativo es el estándar CarbonFix, que permite la comercialización de créditos *ex-ante*, es decir, antes de que se produzca la captura efectiva de carbono (6). Para reducir el riesgo de incumplimiento o de pérdida del carbono almacenado, los estándares de certificación imponen reglas de reserva (*buffer*) y mecanismos de seguro: una parte de los créditos generados se retiene como garantía, de modo que pueda compensarse cualquier liberación posterior de carbono, como la causada por incendios, tala o degradación del ecosistema²⁰². Este tipo de medidas revela que los VCCs son tratados en la práctica como activos digitales transferibles, aunque sin un marco uniforme, como el propuesto por UNIDROIT, que veremos más adelante, puede ocasionar riesgos.

Como otro ejemplo, el mercado voluntario brasileño se estructura casi por completo en torno a estándares privados internacionales. Más del 90% de los créditos emitidos en Brasil entre

²⁰² Como se pone de manifiesto en un caso en Sri Lanka (20).

2017 y 2021 fueron certificados bajo el estándar Verra (VCS)²⁰³, mientras que Gold Standard y American Carbon Registry desempeñaron un papel marginal (8). Esta dependencia refuerza la idea de que, en ausencia de regulación pública, la legitimidad de los VCCs descansa en mecanismos de certificación privados, que actúan como garantes de adicionalidad, permanencia y verificación.

Activos digitales y UNIDROIT: aportaciones a la gobernanza jurídica

El despliegue de los mercados de carbono se inserta en una gobernanza climática crecientemente transnacional, afectada por la diversidad de ordenamientos internos y la ausencia de estándares internacionales homogéneos —incluida la definición de criterios de certificación—. Esta combinación explica tanto los avances como las fricciones del mercado, e impone la necesidad de articular principios comunes que aseguren integridad y eficacia ambiental en el intercambio de créditos (4). El balance es ambivalente: los mercados han sido exitosos financieramente para determinados sectores, pero su impacto agregado no ha logrado frenar el aumento global de las emisiones de GEI (UNEP, 2023) (10). La conclusión operativa es clara: urge homogeneizar parámetros internacionales de certificación y reforzar mecanismos de transferencia hacia comunidades vulnerables, para alinear eficiencia económica con eficacia climática y justicia (4).

El objetivo principal del Proyecto UNIDROIT sobre la Naturaleza Jurídica de los Créditos de Carbono Verificados (Proyecto VCC) es proporcionar orientación sobre cuestiones de Derecho privado, en particular contratos, propiedad, garantías y obligaciones transfronterizas, con el fin de aumentar la confianza en las transacciones de VCCs. Asimismo, busca apoyar el desarrollo de un mercado que funcione con garantías y que pueda desempeñar un papel central en la lucha contra el cambio climático, en el logro de los objetivos de la CMNUCC, en particular el Acuerdo de París, y en la facilitación del cumplimiento de los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas (11). El debate sobre la naturaleza jurídica de los VCCs ha adquirido protagonismo en el marco del Programa de Trabajo 2023-2025 de UNIDROIT, tras la propuesta presentada por la *International Swaps and Derivatives Association* (ISDA) en 2022, que sugería un marco jurídico uniforme para los VCCs.

La evolución de los trabajos de UNIDROIT refleja un tránsito claro desde la fase diagnóstica de 2023-2024 hacia una fase de propuestas en 2025. Mientras los primeros debates subrayaban la fragmentación de registros, la inseguridad jurídica derivada de concepciones divergentes y

²⁰³ Verra es un estándar internacional de carácter voluntario que certifica proyectos destinados a la reducción, eliminación o captura de emisiones de GEI (17).

la falta de garantías para los inversores, el borrador más reciente propone respuestas a estas dudas. A modo de resumen y sin perjuicio de que se estudie con más detalle a continuación, introduce el principio de interoperabilidad entre registros, de modo que lo esencial no sea la existencia de un sistema único, sino que cada registro asegure una identificación inequívoca del crédito, la posibilidad de transferirlo y la exclusión de control de terceros. Asimismo, califica expresamente a los VCC como activos digitales, armonizando su tratamiento jurídico con los *Principles on Digital Assets and Private Law* ya mencionados. Estas soluciones no eliminan por completo la fragmentación estructural del mercado, pero sientan las bases para una gobernanza más segura y previsible. A su vez, la apelación a la interoperabilidad enlaza de manera natural con los debates tecnológicos sobre el uso de *blockchain* e inteligencia artificial en la verificación y trazabilidad de los créditos de carbono, cuestión que se abordará en el epígrafe siguiente.

Mientras que las dos primeras sesiones de 2023 y 2024 fueron más genéricas, en la tercera sesión del Grupo de Expertos (septiembre 2024), el debate se centró en la caracterización de los VCCs como activos digitales, tomando como referencia los *Principles on Digital Assets and Private Law* (2023). Aunque persistieron discrepancias sobre si debían concebirse como bienes de propiedad o como meros derechos contractuales, fue ganando fuerza la noción de “control” como criterio funcional de titularidad, en línea con el enfoque aplicado a otros activos digitales. El Grupo identificó además los objetivos centrales del futuro instrumento: establecer reglas claras de transferibilidad, reconocer la posibilidad de constituir garantías, proteger a los adquirentes de buena fe y reforzar la interoperabilidad entre registros. De manera significativa, se reafirmó que el proyecto debía ser útil para los países en desarrollo, subrayando que, sin un marco uniforme de Derecho privado, la expansión de los VCCs podría profundizar las asimetrías entre Norte y Sur Global.

Esta sesión supuso un punto de inflexión en el proceso de trabajo de UNIDROIT, pues dio lugar a la publicación del primer borrador de Principios. Con este documento, el Grupo de Expertos pasó de los diagnósticos iniciales a la formulación de reglas en forma de *black-letter rules*. A partir de entonces, los trabajos de 2025 se centraron en introducir modificaciones y mejoras sobre ese texto preliminar: mayor precisión en la definición de VCC, aclaraciones sobre revocación y reversión, distinción entre registro, emisión y custodia, y la incorporación de salvaguardas sociales y ambientales para asegurar la coherencia con los ODS. De este modo, el proyecto entró en una fase de consolidación, en la que las propuestas se afinaban hacia un conjunto más estable de principios de Derecho privado aplicables a los créditos de carbono verificados.

En la cuarta sesión (abril 2025), el Grupo de Expertos presentó una versión revisada del borrador inicial, incorporando modificaciones sustantivas destinadas a reforzar la seguridad jurídica de las transacciones con VCCs. Entre las novedades más destacadas figura la definición

funcional de los créditos de carbono verificados como unidades intangibles registradas electrónicamente que representan reducciones de emisiones certificadas por terceros independientes. El texto precisó también los supuestos de revocación y reversión, delimitando las responsabilidades en caso de que la reducción de emisiones no se mantuviera en el tiempo. Asimismo, se profundizó en la distinción entre registro, emisión y custodia de los créditos, con el objetivo de clarificar los distintos roles de los operadores en el mercado. Finalmente, el borrador incorporó referencias explícitas a salvaguardas sociales y ambientales, subrayando que la circulación de VCCs debe realizarse de manera coherente con el desarrollo sostenible y con el respeto a los derechos de las comunidades locales en los países de acogida. Estos avances reflejan el tránsito del proyecto hacia un marco más consolidado, armonizando los VCCs con la categoría de activos digitales definida en los *Principles on Digital Assets and Private Law* y estableciendo principios claros sobre transferibilidad, control, garantías y protección de adquirentes de buena fe.

El documento de abril de 2025 del Grupo de Trabajo consolidó las principales cuestiones que afectan a la circulación de VCC: su calificación jurídica, la necesidad de garantizar su transferibilidad y control exclusivo, la interoperabilidad entre diferentes registros y estándares privados, y la oponibilidad transnacional de las operaciones²⁰⁴.

Tras la revisión de abril de 2025, que consolidó aspectos clave como la definición funcional de los VCCs, la introducción de reglas sobre revocación y reversión y la incorporación de salvaguardas sociales y ambientales, el Grupo de Expertos se reunió nuevamente en Roma en septiembre de 2025 para trabajar sobre una versión aún más detallada del borrador. Esta sexta sesión no supuso un cambio de planteamiento, sino un ejercicio de perfeccionamiento técnico: se precisaron reglas sobre adquisición de buena fe, se afinaron los mecanismos de reversión, se clarificó el papel de los registros —incluyendo la hipótesis de su funcionamiento sobre *blockchain*— y se completaron los capítulos dedicados a custodia, garantías reales e insolvencia. Así, el proyecto entró en una fase de maduración *normativa*, en la que las reglas iniciales fueron transformándose en un cuerpo coherente de principios y comentarios orientados a la práctica del mercado.

El último borrador debatido en Roma (10–12 septiembre 2025) consolida la estructura del instrumento en nueve secciones: definiciones; DIPr; creación/transferencia; cancelación —con subreglas de reversión, revocación y retiro—; registro; custodia; garantías reales; procesal/ejecución; e insolvencia. Los comentarios incluidos precisan la aplicación práctica a lo

²⁰⁴ En concreto, se destacó que los VCCs no pueden entenderse únicamente como instrumentos ambientales, sino como bienes con autonomía patrimonial que participan en el tráfico mercantil, susceptibles de negociación, cesión y gravamen.

largo de todo el ciclo de vida del crédito (generación, negociación secundaria, uso/retiro y cancelación). Este nivel de detalle no existía en versiones anteriores y marca un salto hacia unos principios más sólidos que pueden tomarse como referencia por legisladores y operadores de mercado. Por ser la versión más actualizada hasta hoy, destacamos sus ideas principales que podemos resumir en seis.

I. Adquisición de buena fe y seguridad del tráfico.

El Principio 7²⁰⁵ se refuerza para proteger al adquirente de buena fe (*innocent acquirer*) bajo tres condiciones acumulativas: (a) que el transferente figure como “titular registrado” en el registro pertinente; (b) que el crédito quede efectivamente abonado en la cuenta del adquirente; y (c) que el adquirente cumpla los requisitos de buena fe o “a título oneroso” que disponga el Derecho nacional aplicable. Se aclara expresamente que el registro no garantiza la titularidad, sino actúa como infraestructura neutral con función de publicidad y anotación; la calificación final de la adquisición como “inocente” es cuestión de Derecho interno. Se incorpora, además, una excepción (*carve-out*) para las garantías: el uso del VCC como colateral puede considerarse transmisión con “valor” a efectos de protección del adquirente, aunque no haya un pago inmediato (p. ej., en un préstamo garantizado).

II. Cancelación por “reversión” (reversal).

El Principio 9²⁰⁶ establece que la reversión —pérdida sobrevenida de los beneficios ambientales— solo surte efectos cuando el registro practica el asiento de cancelación, no cuando ocurre el hecho material. Además, introduce un “waterfall” de prioridades para absorber la pérdida: (1) acuerdos válidos entre titulares y el operador del registro que reordenen la prelación; (2) prioridades fijadas por el promotor del proyecto con aprobación del CCB en la emisión; y (3) un régimen supletorio (prorrata, o alternativas FIFO/LIFO que el

²⁰⁵ Principle 7 “Innocent acquisition (1) An innocent acquirer takes a VCC free of conflicting proprietary rights. (2) No rights based on a proprietary claim relating to a VCC can be successfully asserted against an innocent acquirer of that VCC. (3) An innocent acquirer can acquire a proprietary right in a VCC even if the transferor is acting wrongfully and has no proprietary right in the VCC. (4) In order to qualify as an innocent acquirer, a transferee must: (a) acquire the VCC from a transferor that is the registered holder of the VCC at the time of transfer; (b) have the VCC credited to their registry account by the VCC registry operator; and (c) comply with requirements equivalent to those found in the relevant good faith acquisition and take-free rules as specified by the relevant State. (5) A transferee of a VCC is not an innocent acquirer if the transfer of the VCC is (a) made by way of gift or otherwise gratuitously; and (b) is not the grant of a security right.”

²⁰⁶ Principle 9 “Reversal (1) A VCC can be cancelled where it no longer meets the definition of a VCC because of a reversal. (2) A VCC can be cancelled for reversal by a CCB in accordance with the rules of the relevant ICCP, by a court, or by a regulatory body determined by other law. (3) A cancellation for reversal is effective at the time the VCC registry makes an entry indicating its cancellation as such. (4) Although a VCC ceases to be the subject of proprietary rights after cancellation for reversal, cancellation for reversal does not affect such rights that form part of the VCC that its holder may have against any parties (if any). (5) Any cancellation for reversal should be effected in the following order of priority: (a) in accordance with any valid legal agreements between two or more registered holders and the registry operator intended to supersede the priorities in paragraph (b) or (c); (b) in accordance with any priorities established by a project proponent with the approval of a CCB in relation to the particular carbon mitigation project at the time of issuance; or (c) on a pro rata basis among all registered holders of the VCC of the particular carbon mitigation project. Commentary 9.1 [alt1 (c)] on a first-in, first-out basis among all registered holders of the VCC of the particular carbon mitigation project. [alt2 (c)] on a last-in, first-out basis among all registered holders of the VCC of the particular carbon mitigation project.”

texto deja como brackets para decisión final). El comentario subraya que el esquema es no prescriptivo: busca activar soluciones de mercado (seguros, derechos de reposición, compensaciones) y recordar a los agentes que la reversión es un riesgo que debe internalizarse y pre-tasarse en el precio del VCC.

III. Cancelación: reglas comunes y fraccionalidad.

El Principio 8²⁰⁷ organiza los tres supuestos que conducen a cancelación: reversión, revocación y retiro, y prohíbe la cancelación parcial. Si un VCC deja de equivaler una tonelada de CO₂, debe extinguirse íntegramente. Además, aclara las consecuencias registrales: una vez cancelado, el VCC no puede trasladarse entre cuentas ni “retirarse” nuevamente, y si el registro conserva el asiento, debe marcarse el estado de “cancelado” y, de ser necesario, reflejar su retroactividad.

IV. Cancelación por “revocación” (revocation).

El Principio 10²⁰⁸ precisa que, si se demuestra que el crédito nunca cumplió la definición de VCC, la cancelación es retroactiva: el VCC es nulo *ab initio*. No obstante, sobreviven las acciones personales o derechos contractuales que integraran el VCC (por ejemplo, derechos de reposición o compensación monetaria frente a promotor o verificador), preservando remedios frente a terceros distintos del titular.

V. Registro y (hipótesis de) *blockchain*.

Los Principios 12–13²⁰⁹ definen el registro VCC como un sistema operado por una persona jurídica (*operator*), separando funciones de registro y programa de acreditación. En caso de

²⁰⁷ Principle 8 “Cancellation (1) A VCC that is cancelled ceases to be capable of being the subject of a proprietary right even if the VCC registry retains the record of its existence after its cancellation. (2) A VCC can be cancelled as a result of a reversal in a carbon mitigation project, because it has been revoked, or when it is retired. (3) A VCC can only be cancelled wholly and not in part. (4) Upon cancellation, (a) the VCC registry must not comply with any instruction given by the registered holder or a user authorised by the registered holder to move or cancel for retirement the cancelled VCC. (b) the VCC registry no longer owes the VCC’s registered holder the duty set out in Principle 13(1)(d). (c) the VCC registry must record that the VCC is cancelled if it retains a record of its existence. (5) The effective time of cancellation of a VCC will depend on the cause for cancellation. Commentary 8.1 (6) Should a VCC be cancelled in error, the rights and remedies of the parties affected, including the possibility of the restoration of the VCC, is a matter for other law.”

²⁰⁸ Principle 10 “Revocation (1) A VCC can be cancelled where it is subsequently demonstrated that it never met the definition of a VCC and is revoked. (2) A VCC can be cancelled for revocation by a CCB in accordance with the rules of the relevant ICCP, by a court, or by a regulatory body determined by other law. (3) [(Alt Subject to Principle 10(7).] A VCC that has been revoked is void from the outset and never existed as the subject of proprietary rights. (4) Although a VCC was never the subject of proprietary rights once it has been cancelled for revocation, cancellation for revocation does not affect such rights that form part of the VCC that its holder may have against any parties (if any). (5) When a VCC registry cancels a VCC for revocation, it must make an entry indicating that the cancellation is retrospective to the time of creation. (6) Any cancellation for reversal should be effected in the following order of priority: (a) in accordance with any valid legal agreements between two or more registered holders and the registry operator intended to supersede the priorities in paragraph (b) or (c); (b) in accordance with any priorities established by a project proponent with the approval of a CCB in relation to the particular carbon mitigation project at the time of issuance; or (c) on a pro rata basis among all registered holders of the VCC of the particular carbon mitigation project. [alt1 (c)] on a first-in, first-out basis among all registered holders of the VCC of the particular carbon mitigation project. [alt2 (c)] on a last-in, first-out basis among all registered holders of the VCC of the particular carbon mitigation project. [Alt (7)] Where a VCC is subsequently demonstrated that it never met the definition of a VCC as a result of an ICCP’s abandonment of or changes to its relevant methodology in response to evolving scientific knowledge, cancellation for revocation is only effective at the time the VCC registry makes an entry indicating its cancellation as such.”

²⁰⁹ Principle 12 “VCC Registry: Definitions (1) ‘VCC registry’ means an electronic database operated by a registry operator in which the following information is recorded: (a) The unique identifier of a VCC or a block of VCCs; (b) The identifier of the person to whose account the VCC or the block of VCCs is credited. (1 bis) [It must be possible for any person to ascertain from the VCC registry (a) Whether a VCC (or a block of VCCs) is credited to an account in that registry (b) Whether a VCC (or a block of VCCs) is identified by a unique identifier.] (2) ‘Registry operator’ means

utilizar *blockchain*, debe existir un operador con control último del sistema y de los datos. Asimismo, se fijan deberes mínimos del operador: asignar identificador único por VCC (o “bloque”), mantener un solo asiento por crédito, cumplir las reglas del registro y reflejar correctamente las cancelaciones y revocaciones.

VI. Custodia e insolvencia.

Se profundiza el régimen de custodia (Principios 14–17²¹⁰), estableciendo deberes del custodio frente al cliente, protección del “cliente inocente” y tratamiento de los VCCs custodiados en

a legal person who operates a registry. (3) ‘Registry account’ means an account maintained by a VCC registry to which VCCs may be credited or debited. Commentary 12.1 (4) ‘Registry account holder’ means a person who has entered into a registry account agreement with the registry operator in relation to one or more accounts. (5) ‘Registry account agreement’ means the agreement governing the registry account between a registered account holder and the registrar. (6) ‘Registered holder’ means a registry account holder to whose registry account a VCC is credited.” Principle 13 “VCC registry (1) In relation to a VCC credited to a registered holder’s account, a registry operator owes the following duties to the registered holder: (a) An obligation to comply with the rules of the registry and the CCB; (b) An obligation to allocate a unique identifier to the VCC (or to the block of VCCs containing that VCC) and to maintain one, and only one, registry entry in relation to that VCC; (c) An obligation to keep an adequate record of the following matters: (i) the creation of the VCC; (ii) (iii) any credits and debits in registry accounts in relation to the VCC; any retirement of the VCC; (iv) any cancellation of the VCC. (d) An obligation to comply with any instruction given by a registered holder or a user authorised by the registered holder in relation to the credited VCC, unless: (i) the registrar is prohibited from complying with the instruction by other law [including regulatory law], the rules of the relevant CCB or by any agreement between the registry operator and a third party to which the registered holder is a party or has consented; (ii) (i) the registry operator is not obliged, by other law or by the registry account agreement with the registered holder, under certain circumstances, to comply with the instruction. (e) An instruction mentioned in paragraph (d) includes the following: An instruction to move the credited VCC to the account of another person by debiting it from the account of the registered holder and crediting it to the account of that other person; An instruction to cancel the credited VCC for retirement. (2) (a) Subject to subparagraphs (b) and (c), a registry operator has no proprietary right in a VCC credited to an account of a registry account holder in the registry it operates. (b) A registry operator may maintain an account in its own name to which VCCs may be credited and debited. (c) A registry account holder may grant a security right in favour of a registry operator over VCCs credited to the account of that registry account holder. (3) Subject to subparagraph (2)(b) and 2(c), a VCC registered in a VCC registry is not available for the satisfaction of claims of creditors of the registry operator. (4) Subject to subparagraph 2(b) and 2(c), if a registry operator of a VCC registry enters into an insolvency-related proceeding, a VCC registered in that registry does not form part of that registry operator’s assets available for distribution to its creditors. UNIDROIT 2025 – Study LXXXVI – W.G.6 – Doc. 3 53. (5) A registry operator must have a Recovery and Orderly Dissolution Plan providing for preservation of all entries on the VCC registry it operates if the registry operator enters into an insolvency-related proceeding. (6) A registry operator must comply with any instruction from [an ICCP in accordance with its rules][a CCB in accordance with the rules of the relevant ICCP], [by a court, or by a regulatory body determined by other law] to cancel a VCC for reversal in accordance with Principle 9 Or to cancel a VCC for revocation in accordance with Principle 10. (7) A registry operator must (a) record the cancellation of a VCC; (b) not permit any movement by debiting and crediting or cancellation of a cancelled VCC.”

²¹⁰ Principle 14 “Custody (1) ‘Intermediary’ means a person who provides services to another person in respect of a VCC. (2) ‘Client’ means a person to whom an intermediary provides services. (3) ‘Custodian’ means an intermediary who is a registered holder of a VCC and who provides services to another person pursuant to a custody agreement in respect of that VCC. (4) ‘Sub-custodian’ means a custodian who provides services to another custodian pursuant to a custody agreement in respect of that VCC. (5) A custodian ‘maintains’ a VCC for a client if: (a) That custodian is the registered holder of the VCC; or (b) That custodian enters into a custody agreement with a sub-custodian with respect to the VCC [in the circumstances set out in Principle 15(4)]. (6) An agreement between an intermediary and a client is a ‘custody agreement’ if: (a) It relates to a VCC; (b) It is entered into in the course of the intermediary’s ordinary course of business; and (c) The intermediary is obliged to: (i) maintain the VCC for the client; and (ii) [instruct the registry operator or, in the case of sub-custody, the intermediary’s own custodian, to cancel the VCC for retirement if instructed by the client to do so.].” Principle 15 “Duties owed by a custodian to its client (1) A custodian owes the following duties to its client in relation to a VCC that it maintains for that client: (a) the custodian is obliged to safeguard the VCC; (b) the custodian is not authorised to instruct the registry operator to move the VCC to the account of another person, [or use it for its own benefit,] except to the extent permitted by the client and by other law; (c) the custodian is obliged to comply with an instruction given by the client to instruct the registry operator to move the VCC to the account of another person, unless: (i) (ii) the custodian is prohibited from complying with the instruction by other law or by any agreement between the custodian and a third party to which the client is a party or has consented; the custodian is not obliged, by other law or by an agreement with the client, under certain circumstances, to comply with the instruction. (2) Unless prohibited by the custody agreement or by other law, a custodian may maintain VCCs [of the same description] for two or more of its clients as an undivided pool. (3) The duties owed by a custodian to its client may include: (a) the duty to keep a record of VCCs it maintains for each of its clients; (b) the duty at all times to securely and effectively maintain VCCs in accordance with the records it keeps for its clients; (c) [the duty to acquire VCCs promptly if this is necessary to satisfy the duty under subparagraph (b);] (d) the duty to separate the VCCs maintained for clients from the VCCs maintained for its own account. (4) Where authorised by a client or by other law, a custodian may fulfil its duties to its client under this Principle by entering into a custody agreement with a sub-custodian if, under that agreement, the sub-custodian is bound by the duties set out in Principle 15(1) and any other duties imposed on a custodian of VCCs by the applicable Principles law or other private law. (5) A VCC maintained by a

caso de insolvencia del intermediario. Esta regulación se alinea con el enfoque de activos digitales y la experiencia comparada en valores intermediados. La tabla de contenidos y el informe de Secretaría confirman que se han fortalecido también los capítulos de garantías reales (Principios 18–22), incluyendo eficacia *erga omnes* mediante anotación registral o acuerdos de control, prioridades y ejecución, así como la sección de procesal, ejecución e insolvencia.

En comparación con la versión de abril de 2025, el borrador de septiembre introduce una formulación más madura y operativa del régimen propuesto. Refuerza la protección de la titularidad y del tráfico jurídico mediante una regla de adquisición inocente apoyada en la publicidad registral. Asimismo, precisa el sistema de cancelación con reglas temporales y de prelación flexibles, que permiten al mercado internalizar el riesgo a través de mecanismos como los seguros o los derechos de reposición. El texto consolida también un modelo de registro compatible con tecnologías de registro distribuido (DLT), pero preservando siempre la figura de un operador responsable del sistema. Por último, configura una arquitectura normativa más completa en materia de custodia, garantías, ejecución e insolvencia, alineando expresamente los VCCs con la categoría de activos digitales definida en los *Principles on Digital Assets and Private Law* de 2023. Son mejoras incrementales —no un giro de timón—, pero materializan la ambición de dotar a los VCCs de un régimen de Derecho privado previsible e interoperable en mercados internacionales fragmentados. De hecho, la Secretaría compromete la presentación de un borrador completo al Consejo de UNIDROIT en 2026.

custodian for a client may be subject to a security right: (a) granted to that custodian by the client; (b) in favour of that custodian arising by operation of other law; or (c) granted to a third party by the client.” Principle 16 “Innocent client (1) Subject to paragraph 2, if a transferee is a client of a custodian, a VCC is credited to the custodian’s registry account by the VCC registry operator, and the custodian maintains the VCC for the client pursuant to their custody agreement, no rights based on a proprietary claim to that VCC may be successfully asserted against the client. (2) Paragraph (1) does not apply if the client, at the time from which the custodian maintains the VCC for that client, has knowledge that, or ought to know pursuant to the standards applied in the relevant good faith acquisition and take-free rules as specified by the relevant State, that (a) another person has a proprietary right in the VCC; and (b) the transfer violates the rights of that other person in relation to its proprietary right. (3) If VCCs are maintained by a custodian for two or more clients in an undivided pool, paragraphs (1) and (2) apply to each client for whom the VCCs are maintained.” Principle 17 “Insolvency of a custodian and creditor claims (1) A VCC maintained by a custodian for a client is not available for the satisfaction of claims of creditors of the custodian. (2) (a) If a custodian enters into an insolvency-related proceeding, a VCC maintained by a custodian for a client does not form part of that custodian’s assets available for distribution to its creditors. (b) If a custodian maintains a VCC for a client with a sub-custodian, and the custodian enters into an insolvency-related proceeding, the rights it has against the sub-custodian in respect of that VCC do not form part of the custodian’s assets available for distribution to its creditors. (3) If a custodian enters into an insolvency-related proceeding, the insolvency representative must take reasonable steps (a) for a VCC registered in the account of the custodian to be moved to a registry account of the client or of a custodian nominated by that client; (b) for any rights the custodian has against any sub-custodian in respect of a VCC maintained for the custodian’s client to be moved to a registry account of that client or otherwise made accessible to that client, including through moving it to the registry account of another custodian nominated by that client. (4) Paragraphs 5 and 6 apply if all of the following requirements are fulfilled: (a) a custodian enters into an insolvency-related proceeding; (b) VCCs of the same description are maintained by the custodian for two or more clients as an undivided pool; and (c) the quantity of VCCs held by the insolvent custodian for those clients is less than the aggregate quantity of VCCs of the same description that it is obliged to maintain for those clients (‘shortfall’). (5) [The shortfall is met first by any VCCs of the same description maintained by the custodian for itself.] (6) Any [remaining] shortfall shall be borne by the clients for whom the custodian maintains the VCCs as an undivided pool, in proportion to the respective quantity of VCCs of the same description that the custodian is obliged to maintain for those clients. (7) If a custodian maintains a VCC for a client with a sub-custodian and the subcustodian enters into an insolvency-related proceeding, the custodian must seek to cause that VCC to be registered in its own registry account or to maintain the VCC with another sub-custodian.”

Inteligencia artificial y blockchain en la verificación y trazabilidad de los VCCs

La integridad del mercado voluntario depende de la trazabilidad y de la evitación de doble contabilidad. La ausencia de estándares homogéneos de registro y verificación, sumada a asimetrías entre plataformas y a las diferencias en la calidad de los proyectos, erosiona la confianza de inversores y reguladores. De ahí la necesidad de conseguir centralidad en los mecanismos que acrediten el crédito a lo largo de todo su ciclo de vida (5). La tecnología *blockchain* ofrece ventajas claras en este contexto: su registro distribuido y descentralizado, la inmutabilidad de los asientos y el acceso síncrono a la misma información por todas las partes incrementan la transparencia del crédito. Sin embargo, su adopción no sustituye la gobernanza jurídica: la validez del crédito sigue dependiendo de metodologías estandarizadas, verificación independiente y reglas claras de emisión, transferencia y revocación (5)²¹¹. La inteligencia artificial puede ayudar a complementar esta arquitectura al permitir análisis de datos a gran escala; algoritmos de *machine learning* pueden validar patrones de adicionalidad y permanencia, cruzar información de sensores remotos, satélites o informes de campo y emitir alertas tempranas sobre posibles fraudes o errores metodológicos, aunque es algo aún por desarrollar.

Estas propuestas tecnológicas encajan con la aproximación funcional de UNIDROIT: el “control” sobre el activo digital exige una arquitectura que permita excluir a terceros y asegurar la unicidad; la protección del adquirente de buena fe y la oponibilidad jurídica dependen de registros confiables y reglas de cancelación coordinadas. La *tokenización* y la IA solo aportan valor si se integran —contractual y técnicamente— con el registro que confiere existencia, titularidad y unicidad al crédito (5).

En cualquier caso, la falta de reglas internacionales específicas y homogéneas —especialmente sobre certificación, verificación, interoperabilidad y trazabilidad tecnológica— sigue minando la confianza y ha alimentado críticas sobre la eficacia climática global. A pesar de la expansión de estos instrumentos, las emisiones de gases de efecto invernadero continúan aumentando (4) (10), lo que subraya la necesidad de marcos jurídicos y tecnológicos robustos que garanticen un mercado de VCCs de confianza.

Regulación nacional y Derecho comparado en la UE

²¹¹ La experiencia reciente con Verra ilustra el riesgo sistémico de la tokenización desconectada del registrador: se prohibió la emisión de tokens referidos a créditos cancelados o parcialmente consumidos por su potencial de doble contabilidad. La solución propuesta fue la “inmovilización”: bloquear el crédito en el registro mientras se emitían los tokens y, al consumirse, eliminar simultáneamente tokens y crédito. El mensaje es nítido: cualquier esquema crypto debe anclarse al registro primario y a reglas de cancelación sincronizadas (5).

El Reglamento (UE) 2021/1119 traduce en obligación jurídica vinculante el compromiso del Acuerdo de París: alcanzar la neutralidad climática en 2050 mediante el equilibrio entre emisiones y absorciones de GEI, con el objetivo de alcanzar, a partir de esa fecha, un balance negativo de emisiones. Se trata de la pieza central del marco normativo climático de la UE, que dota de seguridad jurídica a las trayectorias de descarbonización de los Estados miembros (12). El Reglamento fija un objetivo intermedio de reducción neta de emisiones de al menos un 55 % en 2030 respecto a 1990²¹². Además del horizonte 2030, el Reglamento impone la obligación de fijar un nuevo objetivo intermedio vinculante para 2040, en coherencia con la senda hacia 2050.

Esta técnica de escalones sucesivos refuerza la previsibilidad en la transición, aunque también genera tensiones políticas y económicas en cada revisión, al exigir mayor ambición respecto de las metas previas. El propio Reglamento reconoce que el cumplimiento de los objetivos no depende solo de la descarbonización del sector energético o del transporte, sino también de la recuperación y gestión de sumideros naturales. La conservación de ecosistemas y biodiversidad se convierte en elemento esencial, lo que refuerza la relevancia de los proyectos forestales y de uso del suelo en los mercados voluntarios de carbono. El Reglamento anuncia la creación de un mecanismo de ajuste en frontera del carbono (CBAM) para determinados sectores, con el fin de evitar que los esfuerzos de la UE se vean neutralizados por importaciones de terceros países con estándares climáticos menos exigentes (12).

En este contexto, debe recordarse que la Unión Europea ha fijado objetivos ambiciosos de reducción de emisiones y ha creado un mercado de derechos de emisión (EU ETS) como eje normativo y económico de su política climática. España, a través de su Estrategia de Transición Justa y del Plan Nacional Integrado de Energía y Clima, se alinea con este marco, que combina normas regulatorias y mecanismos de mercado para alcanzar la neutralidad en 2050 (13).

Los instrumentos de mercado —impuestos ambientales, comercio de derechos de emisión, reformas de subsidios o contratación pública verde— constituyen el núcleo de la transición hacia una economía descarbonizada en la UE. Se trata de aplicar de forma efectiva el principio de «quien contamina paga», internalizando los costes reales de producción y consumo en los precios de mercado y generando incentivos a la innovación y a la eficiencia (14).

²¹² La expresión “al menos” refleja una formulación abierta que abre la puerta a futuros aumentos de ambición, consolidando el principio de irreversibilidad y progresividad en la reducción de emisiones.

En paralelo, la transición hacia una economía baja en carbono y circular implica cambios tecnológicos, organizativos, financieros y sociales a lo largo de todas las cadenas de valor. La UE desarrolla esta visión en su Programa General de Medio Ambiente y en la Estrategia de bioeconomía, vinculando objetivos de descarbonización con políticas de empleo, innovación y cohesión social (14).

Como ejemplo de Estado Miembro que aplica esta estrategia encontramos Portugal, que ha regulado expresamente en su Decreto-Lei n.º 4/2024²¹³ los supuestos de reversión de emisiones, es decir, la pérdida de reducciones previamente certificadas, distinguiendo entre reversiones intencionales y no intencionales. En ambos casos, el promotor está obligado a cancelar créditos equivalentes, pudiendo recurrir a seguros o a una bolsa de garantía alimentada con un porcentaje de los créditos emitidos. Este mecanismo nacional guarda estrecha sintonía con las propuestas de UNIDROIT sobre cancelación y reversión, ofreciendo un ejemplo práctico de cómo internalizar jurídicamente el riesgo de pérdida ambiental (15).

La experiencia portuguesa pone de relieve que los mercados voluntarios nacionales afrontan desafíos similares a los de otros Estados miembros de la UE. La ausencia de metodologías aprobadas sobre neutralidad climática genera incertidumbre, mientras que la creciente presión normativa derivada del Pacto Ecológico Europeo y de instrumentos como el EU ETS o el futuro *Carbon Removal Certification Framework* obligan a los ordenamientos nacionales a adaptarse con rapidez (15).

En el caso español, nuestra jurisprudencia también ha abordado cuestiones relacionadas con los derechos de emisión y la configuración del mercado de carbono, por ejemplo, en la STS 2685/2021, de 14 de junio (Sala de lo Contencioso-Administrativo, Sección 3ª, recurso 220/2020). El Tribunal Supremo analizó el régimen retributivo de las instalaciones de cogeneración (plantas que producen simultáneamente electricidad y calor útil) y observó que éste no preveía la revisión periódica del coste de los derechos de emisión de CO₂, calculado sobre precios históricos. La Sala considera que, aunque sería deseable una metodología más dinámica y transparente, corresponde al *Legislador* y no al juez determinar estos parámetros²¹⁴.

²¹³ *Decreto-Lei n.º 4/2024 | DR.*

²¹⁴ El fallo declara la nulidad de los parámetros retributivos aplicables a las plantas de cogeneración que venden calor a industrias incluidas en el régimen europeo de comercio de emisiones, cuando se computan como coste propio derechos de emisión gratuitos que en realidad perciben las industrias consumidoras. Esta práctica generaba una reducción indebida de la retribución de las instalaciones de cogeneración, contrariando el principio de suficiencia retributiva. El voto particular disiente de la mayoría y defiende que la Orden TED/171/2020 es contraria a Derecho al fijar un valor rígido para el coste de los derechos de emisión de CO₂ basado en datos del pasado. A su juicio, la falta de actualización

Por otro lado, también encontramos la STC 62/2018, de 7 de junio, en la que el Tribunal Constitucional recuerda que cada crédito representa una tonelada de CO₂ reducida o evitada, con precio prefijado, y se adquiere mediante contrato. Esta calificación resalta la dimensión patrimonial y mercantil de los créditos de carbono, diferenciándolos de las ayudas públicas. La sentencia insiste en que el mercado de derechos de emisión exige coherencia y uniformidad para evitar distorsiones competitivas. Una fragmentación autonómica en la evaluación o adquisición de créditos habría generado 17 criterios distintos, quebrando la unidad de mercado. El Tribunal subraya que este principio de uniformidad justifica, en ocasiones, una centralización excepcional de funciones ejecutivas en el Estado.

Derechos humanos, justicia climática y Sur Global

La transición energética no es solo tecnológica, sino también social. El CES europeo destaca que debe garantizarse una transición justa, con apoyo a los trabajadores y territorios más afectados, evitando que la descarbonización se traduzca en pérdida de empleo o desigualdad social. Los instrumentos de mercado, incluidos los créditos de carbono, deben insertarse en esta lógica de equidad para mantener legitimidad social y política (13). De hecho, como señala UNIDROIT, dado que una parte significativa de los proyectos que generan VCCs se localizan en economías en desarrollo, un mercado de créditos de carbono fiable ofrece la oportunidad de canalizar capital hacia los mercados emergentes y financiar proyectos de mitigación climática.

En el caso brasileño, tras la ratificación del Protocolo de Kioto la obligación de restaurar las reservas forestales legales en Brasil se convirtió en una actividad económicamente atractiva, dado el elevado potencial de *secuestro* de carbono que ofrecen los bosques en proceso de crecimiento frente a aquellos ya maduros, siendo estas áreas consideradas potenciales para alcanzar los objetivos comprometidos. La incorporación al Protocolo actuó como catalizador, impulsando sobre todo a los sectores energéticos a promover iniciativas de reducción de emisiones y favoreciendo la creación del Mercado Brasileño de Reducciones de Emisiones (3). Este mecanismo genera un sistema de intercambio de derechos de emisión en el que participan tanto Estados como empresas, y en el que se combinan transacciones en mercados organizados (por ejemplo, bolsas de carbono) y en mercados extrabursátiles mediante contratos. Además, la evidencia empírica muestra que los precios de los créditos voluntarios no son uniformes, sino que varían en función del volumen adquirido y de la posición negociadora de los compradores, situándose en un rango de entre 4 y 10 dólares por tonelada (6). En el caso brasileño, el

perpetúa una insuficiencia retributiva, lo que vulnera el art. 14 de la Ley del Sector Eléctrico. Propone, en línea con la CNMC, introducir mecanismos de revisión o fórmulas que contemplen la evolución futura de los precios.

mercado voluntario de carbono adquirió fuerza a partir de 2019, en paralelo a la incertidumbre sobre la continuidad del MDL. Las emisiones de créditos voluntarios en Brasil crecieron un 779% entre 2019 y 2021, impulsadas por la alta demanda empresarial para cumplir compromisos de neutralidad climática. Este crecimiento refleja cómo el mercado voluntario se ha consolidado como el principal canal de generación y comercialización de créditos en el país (8) (15).

La posibilidad de comerciar créditos de carbono a través de estas distintas plataformas refleja su asimilación jurídica a otros bienes intangibles de carácter mercantil, que pueden circular con relativa facilidad en el tráfico jurídico. Asimismo, se ha desarrollado una dimensión internacional en la que los países industrializados no solo comercian entre sí, sino que también financian proyectos de reducción de emisiones en países en desarrollo mediante la adquisición de créditos de carbono generados en estos territorios. Este diseño ha permitido movilizar capital hacia el Sur Global y, a la vez, ofrecer a los Estados industrializados vías más costo-eficientes de cumplimiento.

Los proyectos voluntarios desarrollados en países como India —centrados en biogás doméstico— o en Uganda —mediante programas de reforestación— muestran el potencial de los VCCs para canalizar recursos hacia comunidades rurales en el Sur Global. Sin embargo, la evidencia también señala que estas iniciativas suelen implicar complejas cadenas de intermediación y costes significativos de certificación, lo que puede reducir el beneficio neto para las comunidades receptoras (6). Sin un marco jurídico uniforme, existe el riesgo de que el mercado de carbono reproduzca desigualdades preexistentes en lugar de corregirlas.

Algunos proyectos voluntarios tienen beneficios “intrínsecos” para los hogares participantes, como ocurre con los sistemas de biogás, las cocinas limpias o los programas de forestación comunitaria. Estos proyectos no solo generan créditos de carbono, sino que también mejoran directamente el acceso a energía, reducen costes domésticos y fortalecen asociaciones locales. Pese a ello, la mayoría de los proyectos de carbono de gran escala no buscan prioritariamente impactos en los medios de vida, sino más bien en el crecimiento económico y la transferencia tecnológica (7).

Al mismo tiempo, los proyectos de carbono pueden generar riesgos para las comunidades locales. Entre ellos se han documentado la captura de beneficios por élites locales, conflictos durante los procesos de titulación de tierras, exclusión de hogares sin títulos claros y restricciones a la gestión comunitaria de los recursos. Estas tensiones ponen de manifiesto que la seguridad jurídica y la gobernanza inclusiva son esenciales para evitar que los mercados de carbono profundicen las desigualdades (7) (16).

En Brasil, aunque la mayoría de los proyectos registrados corresponden al sector energético, la gran mayoría de créditos generados provienen especialmente proyectos REDD+ en la Amazonía (8). Estos proyectos no solo buscan reducir emisiones evitando la deforestación, sino que

también prometen beneficios sociales y económicos para comunidades locales. Sin embargo, la validación y certificación sigue en manos de entidades extranjeras y los costes de transacción pueden absorber gran parte del valor de los créditos, lo que limita el impacto directo sobre los medios de vida en el territorio.

Brasil articula una arquitectura normativa que combina la Política Nacional de Cambio Climático²¹⁵, la previsión de un Mercado Brasileño de Reducción de Emisiones (MBRE) y los contratos de pago por servicios ambientales con carácter *propter rem*, inscribibles en el registro de la propiedad²¹⁶. Este diseño busca otorgar seguridad jurídica real sobre proyectos que generan créditos y anclar su oponibilidad frente a terceros (4). Aunque Brasil ha institucionalizado REDD+ y avanza con otros instrumentos, persiste una dispersión normativa entre niveles y sectores. La literatura resalta la necesidad de una codificación moderna y unificada para que el país pueda realizar su potencial de oferta en el mercado global sin sacrificar integridad ni seguridad jurídica (4).

Como se ha comentado previamente, los mercados de carbono solo tendrán legitimidad si se insertan en una transición justa: mientras algunos sectores pierden empleo (carbón, industrias intensivas en energía), otros lo ganan (renovables, eficiencia energética, rehabilitación)²¹⁷. El balance neto es positivo, pero requiere políticas de apoyo, recualificación y protección social que aseguren que los beneficios de la descarbonización se distribuyan de manera equitativa (14).

Conclusiones

El presente trabajo ha examinado la naturaleza jurídica de los créditos de carbono voluntarios (VCCs) y su inserción en la arquitectura global de gobernanza climática. A partir del análisis de los mecanismos del Protocolo de Kioto y del Acuerdo de París, se ha demostrado que los mercados regulados de carbono sentaron las bases conceptuales de la mercantilización de reducciones de emisiones. Sin embargo, han sido los mercados voluntarios los que, en ausencia de obligaciones jurídicas internacionales, se multiplicaron bajo estándares privados y autorregulados. Este crecimiento desordenado, aunque permitió canalizar recursos hacia

²¹⁵ Ley 12.187/2009, [L12187](#).

²¹⁶ Ley 14.119/2021, [L14119](#).

²¹⁷ Otro riesgo señalado es el del *greenwashing* laboral o simulación de empleos verdes, es decir, que la transición cree empleos verdes precarios o falsos autónomos, reproduciendo desigualdades en lugar de corregirlas. Así como se exige integridad ambiental a los créditos de carbono, también debería reclamarse integridad social, garantizando que los proyectos certificados no solo cumplan con la integridad ambiental, sino también con la integridad social (14).

proyectos de mitigación, también generó un mosaico de reglas heterogéneas y de proyectos diversos.

Desde un punto de vista jurídico, la principal aportación de este estudio es analizar los VCCs como activos digitales intangibles, siguiendo la aproximación funcional de los *Principles on Digital Assets and Private Law* de UNIDROIT. Este enfoque, centrado en el “control” como criterio de titularidad y transferibilidad, ofrece un marco más coherente que las concepciones tradicionales basadas en derechos personales o meramente contractuales. La equiparación de los VCCs con otros activos digitales permite, además, incorporar instituciones ya asentadas en el Derecho privado: la protección del adquirente de buena fe, la constitución de garantías reales sobre bienes intangibles y la interoperabilidad entre registros electrónicos.

El análisis de la experiencia normativa europea, la legislación climática confirma que los instrumentos de mercado son piezas esenciales de la transición ecológica, pero también muestra que la coherencia interna y la uniformidad son condiciones indispensables para evitar distorsiones entre los diferentes Estados Miembros.

A nivel doctrinal, se ha constatado que la inseguridad jurídica de los VCCs no puede resolverse únicamente con mejoras en los estándares privados de certificación. Es imprescindible un marco uniforme de Derecho privado transnacional, capaz de reconocer a los VCCs como activos digitales, establecer reglas claras sobre su emisión, transferencia, cancelación y custodia, y garantizar su oponibilidad frente a terceros. En este sentido, el Proyecto de UNIDROIT sobre Créditos de Carbono puede llegar a consolidarse como un avance decisivo, al proporcionar principios claros sobre la adquisición de buena fe, la interoperabilidad entre registros o el tratamiento de la insolvencia.

El trabajo también ha resaltado la dimensión de justicia climática. La mayor parte de los proyectos que generan VCCs se desarrollan en el Sur Global, en territorios donde la debilidad institucional y la falta de recursos aumentan la vulnerabilidad frente a prácticas abusivas. Aunque los proyectos voluntarios pueden aportar beneficios intrínsecos a las comunidades — acceso a energía, reforestación comunitaria, reducción de costes domésticos—, la evidencia demuestra que también pueden reproducir desigualdades: apropiación de beneficios por élites locales, conflictos de titulación de tierras, exclusión de hogares vulnerables o empleos precarios bajo apariencia de “empleo verde”. De ahí que el reconocimiento de los VCCs como activos digitales deba acompañarse de salvaguardas sociales efectivas, capaces de garantizar que la integridad ambiental vaya de la mano de la integridad social.

De todo ello se desprende que los VCCs solo alcanzarán legitimidad y eficacia climática si cumplen simultáneamente tres exigencias. En primer lugar, integridad ambiental, es decir, que representen reducciones reales, adicionales, verificadas y permanentes de las emisiones de carbono. En segundo lugar, seguridad jurídica, que garantice su naturaleza como activos

digitales transferibles, cuyos propietarios queden protegidos frente a fraudes. En tercer lugar, un eje de justicia social y climática, que asegure que los beneficios lleguen a las comunidades locales y que la transición no reproduzca desigualdades.

En definitiva, el estudio de los VCCs revela que estamos ante un instrumento en evolución, con un potencial significativo, pero también con incertidumbres notables. La construcción de un régimen de Derecho privado transnacional reforzado por soluciones tecnológicas responsables, constituye una condición necesaria para que los mercados voluntarios de carbono evolucionen de un mosaico fragmentado hacia un sistema uniforme, capaz de alinear eficiencia económica, seguridad jurídica y justicia climática.

Bibliografía

1. *El Protocolo de Kyoto y los Bonos de Carbono*. Anthony Laub Benavides, Javier Matos Flores Guerra. 2008, Revista de Derecho Administrativo, Vol. 6, págs. 239-248.
2. *Bonos de carbono*. Conterol. 4, 2021, Interempresas. Industria química, Vol. 94, págs. 104-105.
3. *O princípio do desenvolvimento sustentável e o mercado de créditos de carbono*. Goncalves, Naila. 2016, Revista: DELOS Desarrollo Local Sostenible, Vol. 16.
4. *Mercado de créditos de carbono na perspectiva da governança climática transnacional: experiências do Brasil e da Colômbia*. Zenildo Bodnar, Javier Gonzaga Valencia Hernández. 3, 2022, Novos Estudos Jurídicos (NEJ), Vol. 27, págs. 452-474.
5. Oliveira, Izadora Gabriele dos Santos. NFTS e sua aplicação ao mercado de créditos de carbono. [aut. libro] VVAA. *Liber Amicorum: "In memoriam" al prof. Dr. Félix Salvador Pérez*. Sevilla : Universidad de Sevilla, 2022, págs. 193-212.
6. Peskett, Leo, Brown, Jessica y Schreckenber, Kate. *CARBON OFFSETS FOR FORESTRY AND BIOENERGY: RESEARCHING OPPORTUNITIES FOR POOR RURAL COMMUNITIES*. Ford Foundation. s.l. : Overseas Development Institute, 2010.
7. Group, The World Bank. *CARBON LIVELIHOODS: Social Opportunities & Risks for Carbon Finance*. 2012.
8. Daniel Barcelos Vargas, Linda Márcia Mendes Delazeri, Vinícius Hector Pires Ferreira. *Voluntary Carbon Market in Brazil in Reality and in Practice*. São Paulo : Escuela de Economía de São Paulo, VSG, 2021.
9. UNEP. *Emissions Gap Report 2023*. Nairobi : s.n., 2023.
10. H. Lee, J. Romero. *Climate Change 2023: Synthesis Report* Core Writing Team. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, IPCC. Ginebra : s.n., 2023.
11. UNIDROIT International Institute for the Unification of Private Law. *VERIFIED CARBON CREDITS*. [En línea] [Citado el: 20 de agosto de 2025.] <https://www.unidroit.org/work-in-progress/verified-carbon-credits/#1631778836762-7d8b7c54-a3cb>.
12. *Reglamento (UE) 2021/1119 del Parlamento Europeo y del Consejo de 30 de junio de 2021 por el que se establece el marco para lograr la neutralidad climática y se modifican los Reglamentos (CE) 401/2009 y (UE) 2018/1999 («Legislación europea sobre el clima»*. García, Sara García. 2021, Actualidad Jurídica Ambiental, Vol. 115, págs. 39-40.

13. *Mercados, entidades financieras y bancos centrales ante el cambio climático: retos y oportunidades*. Clara Isabel González Martínez, Soledad Núñez Ramos. 2019, Presupuesto y gasto público, Vol. 97, págs. 213-234.
14. *La transición hacia una economía baja en carbono: una oportunidad para el clima y el empleo*. Leiva, José Luis de la Cruz. 2, 2016, Administración & ciudadanía: revista da Escola Galega de Administración Pública, Vol. 11, págs. 281-304.
15. *O mercado voluntário de carbono em Portugal*. Gonçalo Andrade Sousa, Francisco Fráguas Mateus. 2024, Actualidad jurídica Uría Menéndez, Vol. 65, págs. 82-92.
16. Barcelos Vargas, Daniel, Mendes Delazeri, Linda Márcia y Pires Ferreira, Vinícius Hector .
17. *Desarrollo sostenible en Guatemala*. Sartori, Angelo. septiembre, 2021, La Revista de la Evaluación de la Conformidad, Vol. 372.
18. *El Protocolo de Kyoto y los Bonos de Carbono*. Anthony Laub Benavides, Javier Matos Flores-Guerra. 2008, Círculo de Derecho Administrativo, Vol. 6, págs. 239-248.
19. *Carbon taxes versus cap and trade: a critical review*. Goulder, Lawrence H., Andrew R. Schein. 03, 2013, Climate Change Economics, Vol. 4.
20. *Livelihood and ecosystem benefits of carbon credits through rainforests: A case study of Hiniduma Bio-link, Sri Lanka*. SENADHEERA, DK Lakmini, WAHALA, WMPSB y WERAGODA, Shermila. 2019, Ecosystem Services, Vol. 37.

ACCEDE A LA GRABACIÓN

Resumen

Desde sus orígenes, que se remontan a la Francia posrevolucionaria, el Derecho Administrativo ha buscado conciliar la garantía de la libertad individual y de los derechos fundamentales con la imposición de límites al ejercicio de la función administrativa por parte del Estado. La inteligencia artificial representa la tecnología más recientemente incorporada al Derecho Administrativo y, dependiendo de su forma de uso, puede dificultar el control de la actividad administrativa y vulnerar derechos y garantías fundamentales. Utilizando el mito de Prometeo, proponemos como solución a este problema jurídico la comprensión de la inteligencia artificial como una tecnología, y no meramente como una técnica. Ante este escenario, se plantea que el Derecho Administrativo conciba la inteligencia artificial como tecnología, entendida como la asociación entre conocimiento y técnica. La tecnología, en la actividad administrativa, se manifiesta en el acto administrativo, concretamente en uno de sus presupuestos de validez: el motivo. Sostenemos que esta concepción jurídica de la inteligencia artificial permitiría el control efectivo de la Administración Pública y la salvaguarda de los derechos fundamentales.

Palabras clave: inteligencia artificial; tecnología; Derecho Administrativo; derechos fundamentales; motivo del acto.

²¹⁸ Este trabajo es una traducción y fue desarrollado durante una estancia de investigación proporcionada por el Departamento de Derecho de la Universidade Federal de Viçosa, en el marco del Grupo de Investigación Regulation, coordinado por los profesores Andrés Boix Palop y Gabriel Doménech Pascual, del Departamento de Derecho Administrativo de la Universitat de València, València, España. La versión completa de esta traducción fue publicada en la Revista de Informação Legislativa: Lelis, D. A. S. de. (2025). *Nem rápido, nem devagar: a natureza jurídica da inteligência artificial no Direito Administrativo europeu continental*. Revista de Informação Legislativa: RIL, 62(247), 37-63. https://doi.org/10.70015/ril_v62_n247_p37

²¹⁹ Doctor en Derecho Público por la Pontificia Universidade Católica de Minas Gerais; Profesor de Derecho Administrativo en el Departamento de Derecho de la Universidade Federal de Viçosa. Correo electrónico: davi.lelis@ufv.br

1. Introducción

La actual revolución tecnológica, liderada por la informática, con la inteligencia artificial (IA), está alterando el papel del Derecho, como si estuviésemos abriendo una caja de Pandora que traerá más problemas que soluciones. Pandora, la primera mujer, fue creada por Hefesto a pedido de Zeus, como una forma de vengarse de la humanidad después de que el titán Prometeo les diera a los hombres el secreto del fuego. Pandora fue enviada a la Tierra para casarse con Epimeteo, hermano de Prometeo, y llevaba consigo una caja que jamás debía ser abierta. Sin embargo, incapaz de contener la curiosidad, Pandora abre la caja y, con ello, libera de su interior todos los males hasta entonces desconocidos por los hombres, como las enfermedades, la guerra, la mentira y el odio. Prometeo, por haber robado el fuego de los dioses, fue encadenado a una roca, condenado a ser devorado vivo todos los días por un águila. De los mitos se extrae que el conocimiento del fuego y la curiosidad transformaron a la humanidad, ahora el conocimiento de la IA y la curiosidad están transformando el Derecho.

Sugerimos un paso en la discusión, para cuestionar si el Derecho ya contempla, en su sistema, la capacidad de responder al avance tecnológico y la implementación de IA en sus actividades, sin vulnerar derechos y garantías fundamentales (LELIS, 2025). Lo que se busca, ante las nuevas aplicaciones de sistemas inteligentes al Derecho, es la adecuación entre tecnología y Derecho, para fines de su aplicación, posibilitando el control de los actos administrativos realizados con la ayuda de la IA. La pregunta que responder es: ¿cuál es la naturaleza jurídica de la IA para el Derecho Administrativo?

La respuesta a esta pregunta ya ha sido ofrecida, con un vocabulario diferente, por Huergo Lora (2020), con la afirmación de que no hay ninguna naturaleza jurídica para los algoritmos predictivos; Boix Palop (2020), al afirmar que los algoritmos son reglamentos; y Berning Pietro (2023), que entiende que debe haber una naturaleza jurídica, pero con la creación de una nueva categoría jurídica para los algoritmos. Lo que se pretende es presentar una propuesta para la naturaleza jurídica de la IA, en la teoría general del Derecho Administrativo, sin la necesidad de ninguna alteración normativa o modificación del Derecho.

Para responder a la pregunta propuesta, presentaremos lo que se entiende por naturaleza jurídica, seguida de una breve definición de tecnología, para comprender qué son los sistemas inteligentes, la IA y los algoritmos; luego presentaremos el debate actual sobre qué es la IA para el Derecho Administrativo; a continuación, demostraremos, considerando una teoría general del acto administrativo, los argumentos que nos llevan a la idea de que la IA es el motivo del acto administrativo y la forma de implementar esta idea para permitir el control de la Administración Pública y la materialización de derechos y garantías fundamentales.

2. Dar nombre a las cosas para entender la naturaleza jurídica, tecnología e IA.

La idea de naturaleza jurídica es tributaria del derecho privado, que, en los años 1900, obedeciendo a la racionalidad de la modernidad, pasó a entender que era necesario atribuir juridicidad a conceptos e institutos, hoy abaricable también a derechos subjetivos. El Derecho moderno, según se enseña en la tradición europea continental, agrupa conceptos emulando sistemas lógicos-matemáticos que parten de una verdad no demostrable, de la cual se deducen conclusiones igualmente verdaderas, respetando principios de no contradicción. Se trata de una racionalidad de generalizar y especificar, transformando el sistema jurídico en un sistema de normas y conceptos, en el cual el contenido de la norma especial se deduce de la norma general, formando, así, institutos jurídicos. Esta práctica de deducción lógica, aunque común, deja escapar otra visión: la de los problemas jurídicos, que se añade a la idea de naturaleza jurídica al establecer que, si un *Hecho X* tiene la misma naturaleza jurídica que el *Hecho Y*, los efectos jurídicos deben ser los mismos.

La naturaleza jurídica no puede ser mera descripción de un hecho de la realidad dentro del sistema jurídico. Se trata de estudiar un fenómeno procurando una clasificación que lo sistematice y lo explique, posibilitando una correcta respuesta jurídica al problema. La naturaleza jurídica es una decisión sobre la normatividad de un concepto, repleta de contornos culturales, políticos, sociales y económicos, con vistas a atender a un propósito. Cuando surge un nuevo derecho, hecho social o económico, para fines de su aplicación jurídica, se busca la naturaleza jurídica, para que se defina el régimen jurídico a ser aplicado. Señalar la naturaleza jurídica no es una excusa para no motivar la decisión, visto que no se trata de una verdad dada sino de una argumentación en el plano de los fines para descubrir a qué objetivo, en un dado contexto, sirve ese derecho o nuevo hecho social o económico. La naturaleza jurídica debe ser útil para la solución de problemas jurídicos.

Si hay un problema en la adopción de la IA por el sector público, en cuanto a la posibilidad de control de la actividad administrativa y no violación de derechos y garantías fundamentales, es preciso saber para qué fin jurídico sirve la tecnología de IA y qué régimen jurídico debe aplicarse a ella. El tema encierra dos expresiones que merecen atención antes de que se pueda efectivamente llegar a la respuesta: cuál es el significado de tecnología y de IA.

Primero la tecnología, expresión polisémica, que ganó importancia en la modernidad, cuando pasa a expresar la idea de competencia de producir. La construcción del concepto de tecnología viene de la Grecia antigua: deriva de la palabra griega *techné*, que dio origen tanto a la palabra técnica, como a la palabra tecnología. Técnica significa fabricar, producir, construir; y se relaciona con la existencia de un instrumento. Tecnología además de *techné*, contiene, también, *logia*, del griego logos, que significa razón; siendo, por lo tanto, etimológicamente, la razón del saber hacer, el estudio de la técnica.

La etimología de la palabra tecnología no es suficiente para la comprensión de su densidad teórica, siendo necesario llegar a una concepción que sea adecuada al actual momento socioeconómico. Paiva (1999) y Veraszto et al (2009) ofrecen un cuadro de las diversas concepciones de tecnología, en la actualidad: (i) la concepción intelectualista comprende la tecnología en sentido práctico y dependiente de la teoría. Así, la ciencia sería jerárquicamente superior a la tecnología; (ii) la concepción utilitarista de la tecnología la toma como sinónimo de técnica; (iii) mientras la concepción científica de la tecnología, la considera como sinónimo de ciencia; (iv) la concepción instrumental de la tecnología la entiende como los instrumentos creados para las diversas tareas enfrentadas por la humanidad, lo que igualaría la piedra lascada de la antigüedad a los actuales utensilios tecnológicos, bastando saber usar el instrumento, sin la necesidad de dominar el proceso y el conocimiento de creación del instrumento, para ser un experto en tecnología; (v) la concepción de neutralidad de la tecnología no la considera ni buena, ni mala, siendo su uso el responsable de los efectos sociales; (vi) la concepción universalista de la tecnología entiende que ella puede surgir en cualquier lugar y ser útil en cualquier contexto y; (vii) la concepción del determinismo tecnológico considera a la tecnología como natural o auto evolutiva.

Ante tantas y diversas concepciones de tecnología, se nota que hay ideas optimistas, que ven en la tecnología la garantía de la supervivencia humana y su dominio sobre la naturaleza; y hay ideas pesimistas, que observan las distorsiones sociales y económicas que la tecnología puede causar. El pesimismo es justificable por las alegorías que el estudio de la técnica y la tecnología comúnmente invocan: el mito de Prometeo y el Fausto, de Goethe (PAIVA, 1999: 7).

Narra la mitología griega, que Prometeo y su hermano Epimeteo, recibieron de Zeus la tarea de atribuir cualidades a la creación. Epimeteo se encargó de la tarea, distribuyendo cualidades variadas a los animales, sin que quedara nada para los humanos. Prometeo, entonces, robó el fuego de los dioses y se lo dio a los humanos, para que pudieran defenderse de los otros animales. Sin embargo, el fuego hizo a los humanos demasiado-poderosos, siendo necesaria la intervención de Zeus, que castigó a Prometeo, encadenándolo a una roca, en el Cáucaso. En el Poema de Goethe, el Doctor Fausto, con deseo de superar el conocimiento de su tiempo, vende su alma al demonio Mefistófeles, que le proporciona una energía transformadora. Al término del contrato, Fausto pierde a su amada Margarita y es llevado al infierno.

Hay tragedia en las dos historias, pero la motivación de los personajes es distinta. Mientras Prometeo tiene afecto por la humanidad y desea ayudarla a transformar y modificar la naturaleza, Fausto tiene interés en superar sus propias dificultades contra la tiranía de la naturaleza, en nombre del libre espíritu que protege sus derechos individuales. La diferencia del motivo apunta a la diferencia entre la técnica y la tecnología: Prometeo es técnica, porque solo pretende dar un instrumento a los hombres, sin reconocer la capacidad transformadora

de este instrumento; mientras que Fausto es tecnología, pues asocia el conocimiento a la técnica.

Tecnología no es una invención, no está en el plano de las ideas, tiene el objetivo de combinar teoría con producción y eficacia. Los conceptos de la ciencia son utilizados por la tecnología, pero solo cuando se reducen a conceptos individuales, posibilitando la construcción y modificación del mundo (ethos), teniendo, por lo tanto, contenido ético y político (PAIVA, 1999: 8). Hay que decir que se hace uso de la tecnología no se limita a la capacidad de desarrollar utensilios, aparatos y herramientas. Hacer uso de la tecnología se extiende a la capacidad de desarrollar tecnologías simbólicas, de la inteligencia, como lenguaje, escritura, arte y Derecho. “Cada sociedad crea, recrea, piensa, repiensa, desea y actúa sobre el mundo a través de la tecnología y de otros sistemas simbólicos. La tecnología es impensable sin admitir la relación entre el hombre y la sociedad” (VERAZTO, 2009: 37).

La tecnología, por lo tanto, sería el conjunto de conocimientos inherentes al desarrollo y concepción de instrumentos (artefactos, procesos, sistemas, lenguaje, etc.), creados por la humanidad, dentro de un dado tiempo histórico, para satisfacer sus intereses personales y colectivos. La tecnología exige un conocimiento del porqué y del cómo se logran sus objetivos, son actividades asociadas a símbolos, instrumentos y máquinas para la creación de obras, productos, teorías y métodos en la ciencia y conocimientos modernos. Y una vez puesta a disposición de la sociedad, del mercado o de otro sistema, como el Derecho, la tecnología pasa a tener el valor determinado por la forma en que va a ser adquirida y usada. Y quien define ese valor es la propia sociedad, el mercado o los expertos en un dado sistema afectado por la tecnología.

Para comprender el significado de IA, comencemos con distinciones: el uso de IA por parte de la Administración Pública no se confunde con la Administración Electrónica, medio por el cual la relación entre Administración y administrado ocurre intermediada por un proceso electrónico. Esa “administración sin papel” (HUERGO LORA, 2020: 26) funciona exactamente como la administración tradicional, con las mismas garantías y presupuestos, no habiendo, en esta aplicación, un problema teórico-jurídico a ser solucionado. En Brasil, por ejemplo, el uso del Sistema Eletrônico de Informações (SEI), regulado por el Decreto nº. 8.539 de 8 de octubre de 2015, no se aparta de las especificaciones de la Constitución, de la Ley nº. 9.784 de 1995, que trata de las normas generales del Proceso Administrativo y de otras normas específicas que por ventura se aplican a determinado procedimiento. En España lo mismo ocurre con las Leyes 39 y 40 de 2015 y con el Decreto Real 1065 de 2015.

La IA trabaja con sistemas inteligentes que, al contrario del mero proceso electrónico, alteran la actuación administrativa. Estos sistemas inteligentes están dotados de algoritmos inteligentes, capaces de sustituir la racionalidad humana. Algoritmo, viene de la forma latina

de escribir el apellido del matemático Abu Jafar Mohamed Ibn Musa Al-Juarismi (780-850), y hace referencia a cualquier procedimiento formalizado en una serie de pasos para solucionar un problema o conseguir un resultado. En sentido amplio, los procedimientos judiciales y administrativos serían algoritmos, ya que son un conjunto de pasos para solucionar un problema. La declaración anual del impuesto sobre la renta, por ejemplo, hecha mediante un programa, es una aplicación algorítmica, en la que el contribuyente solo llena los datos en los lugares adecuados y obtiene el resultado de cuánto debe pagar o cuánto debe recibir de devolución. En estos casos, el algoritmo traduce la norma jurídica creada de acuerdo con el debido proceso legal, a un programa, para facilitar su eficiencia y eficacia. Otros usos de algoritmos en el Derecho incluyen, pero no se agotan en, actualizaciones de valores, revisión de precios y cumplimiento de créditos universitarios y, también, como base para la realización de la administración electrónica.

Otra distinción necesaria es la de separar la IA y los sistemas inteligentes del algoritmo: IA es una expresión creada en 1956 por el matemático John McCarthy, con ocasión de la organización de un grupo de estudios en la facultad de Dartmouth, en el verano de ese año. IA hace referencia a cualquier realización computacional de tareas que anteriormente exigirían la intervención racional humana y tiene por objetivo la construcción de máquinas inteligentes (MITCHELL, 2019: 21). En la definición de Häuselmann (2022: 46), la IA se refiere a máquinas con capacidad de adaptación para autónomamente, ejecutar actividades y tareas que requieren capacidad normalmente asociada con la cognición humana.

Häuselmann (2022: 47) señala seis diferentes sistemas inteligentes: Machine learning, deep learning, natural language processing, emotion AI, computer vision y automated reasoning; y Mammen (2024) añade otro: Federate learning. Hay, por lo tanto, IA capaz de: presentar soluciones basadas en datos preprogramados (Machine learning) o en algoritmos predictivos capaces de aprendizaje a partir de datos brutos, como un texto o imagen, estructurados en redes neuronales artificiales (Deep learning); con la construcción de un modelo en el cual diversas personas pueden participar en la creación y operación del modelo, sin tener contacto con datos que no les pertenecen (federate learning); aprender y procesar el lenguaje humano oral y escrito (Natural language processing), comprendiendo, en el proceso, las variaciones emocionales de quien habla o escribe (Emotion AI); entender imágenes, posibilitando el reconocimiento facial y otras biometrías (Computer vision), pudiendo ser posible, también, el reconocimiento de emociones (Emotion AI); desarrollar ordenadores capaces de almacenar información para responder preguntas en diversos campos del conocimiento, presentando nuevas conclusiones (Automated reasoning). En este trabajo, hay una razón práctica para diferenciar entre los diferentes sistemas inteligentes en dos grupos: los sistemas inteligentes que aprenden, utilizados en actos administrativos discrecionales, y los sistemas inteligentes expertos, sin capacidad de aprendizaje, utilizados en actos administrativos vinculados.

La base de los sistemas inteligentes son los datos que, actualmente, debido a la informatización de (casi) todas las actividades, son abundantes. Sin entrar en los problemas éticos y jurídicos de la recopilación de datos, no basta con la existencia y disponibilidad de esta información para la IA; es necesario extraerla y tabularla para su uso en algoritmos. En algunas áreas, especialmente en el sector privado, el uso de sistemas inteligentes ya ha sustituido el uso de la racionalidad humana. Es lo que ocurre, por ejemplo, con empresas como TripAdvisor, Google Reviews y Google Maps, que usan la sabiduría de las masas, en forma de datos, para hacer recomendaciones a los usuarios, o como Amazon y Spotify, que hacen recomendaciones de productos, basadas en actividades anteriores de sus usuarios (COECKELBERGH, 2023: 23).

Para el Derecho, las aplicaciones de IA también ya presentan usos, como el programa Watson and Debater, de IBM, que analiza textos jurídicos y es capaz de presentar respuestas jurídicas a los problemas propuestos, con base en machine learning. El modelo actúa como un jurista actuaría en busca de la naturaleza jurídica de un hecho de la realidad, correspondiendo al sistema inteligente, ante la situación presentada, decidir cuál es el sentido de la norma jurídica y si se aplica, o no, al caso (ASHLEY, 2017: 28 y ss.). González (2020: 185 y ss.), presenta el VERIPOL, de España, Prepol, HunchLab y Beware, de Estados Unidos, como sistemas inteligentes para la actuación predictiva policial, basada en el análisis de datos que son proporcionados por la denuncia formal (VERIPOL), o por lectura en una base de datos que puede incluir hasta la actividad en redes sociales (Beware). En Brasil el Tribunal de Cuentas de la Unión hace uso del sistema Alice, que lee las licitaciones y pliegos publicados en los Diarios Oficiales, para señalar indicios de fraude (BRASIL, 2021: 41).

Los usos y posibilidades para la aplicación de la IA, en el Derecho Administrativo, son amplios y normalmente son tratados como avances indiscutibles de la eficiencia y de la capacidad humana de dominar la realidad. La IA, para Kleijn (2022), surgió como un intento de replicar, en una máquina, la inteligencia humana; sin embargo, aún no está claro qué camino deben tomar los sistemas inteligentes: el de replicar el cerebro humano o solo sus capacidades cognitivas. Lo cierto es que la IA no es una evolución de la inteligencia humana. Ambas ocurren de formas distintas y presentan desempeños diferentes dependiendo del problema y propósito propuestos.

La inteligencia humana es simbólica, con capacidad de abstracción, metáforas y autorreconocimiento, y se vale de la evaluación de medios y fines para llegar a la respuesta del problema dado, habiendo también interpretación de la realidad. Así, para un humano es trivial la comprensión de problemas que contengan conocimientos de sentido común, mientras que para la IA es necesario proporcionar diversas informaciones complementarias e instrucciones que eviten una interpretación literal del problema. Según Häuselmann (2022: 45), los problemas más difíciles, pero específicos, como jugar al ajedrez con excelencia, resolver ecuaciones de álgebra o de lógica, son los campos en los cuales la IA presenta un resultado

mejor que la inteligencia humana; mientras que los problemas más amplios, como reconocer un objeto y saber que una nube que se parece a un perro no es un perro; exigen más trabajo y datos de los programadores, teniendo la IA resultados peores en comparación con los humanos. Esto se debe a que los humanos aprenden simbólicamente, con prueba y error, realizando diferentes acciones y aprendiendo con los resultados obtenidos; mientras que la IA necesita ser alimentada con datos correctamente preparados para la solución de ese problema específico (KLEIJN, 2022: 33 y ss.).

La IA y sus sistemas inteligentes no son mera técnica de realizar la administración pública. Son tecnologías, conocimientos aplicados a fines, con contenido ético y político, capaces de alterar la forma en que la sociedad, el mercado y la Administración Pública se relacionan. La visión pesimista, del mito prometeico, se justifica porque hay una tendencia de la IA a aumentar los sesgos sociales (MITCHEL, 2019: 119), en especial para con minorías, con violaciones de derechos de igualdad y de no discriminación (SORIANO ARNANZ, 2023). El uso de la IA en la Administración Pública, por lo tanto, exige la cautela que las alegorías de Prometeo, Pandora y Fausto nos advierten. Cada sociedad actúa sobre el mundo a través de la tecnología y de otros sistemas simbólicos. Crear, pensar y actuar sobre el mundo con el Derecho (sistema simbólico) creado por humanos, implica un tipo de racionalidad; crear, pensar y actuar sobre el mundo con el Derecho (sistema matemático) creado por IA implica otro tipo de racionalidad. La IA, en estos términos, no es solo una técnica facilitadora, es una tecnología transformadora de la realidad y de la experiencia humana. Como tecnología, el Derecho tiene la capacidad de juridificar otras tecnologías, como la IA. Es preciso, por eso, dentro del Derecho, en cada rama, pensar qué régimen jurídico aplicar, de forma que se contengan las consecuencias negativas de la tecnología, pues, de lo contrario, sufriremos como Fausto y Prometeo, que desconsideran y no se preocupan por las consecuencias de la opción de adoptar una tecnología en la existencia humana.

3. Nada jurídico o reglamento: El debate actual sobre qué es la IA para el Derecho Administrativo

Entre los académicos, se han formado tres corrientes en cuanto a la comprensión de qué es la IA para el Derecho Administrativo. Boix Palop (2020), pionero en la discusión, opta por buscar la naturaleza jurídica de los algoritmos, mientras que Huergo Lora (2020), niega la existencia de naturaleza jurídica para la IA, demostrando los usos de los algoritmos predictivos en la Administración Pública y Berning Prieto (2023), imagina la posibilidad de crear una nueva categoría jurídica para los algoritmos.

Según Boix Palop (2020), no hay diferencia entre escribir una norma a mano, a máquina o con algoritmos. Ante la posibilidad de la incorporación tecnológica, sería posible que la IA reemplace, en la estructura del Derecho Público, una actividad típicamente humana: la discrecionalidad, y cree regulaciones que regirán la actividad administrativa. Ante el problema

sin precedentes, el Derecho Público, renunciará a algunas de sus estructuras tradicionales en busca de una solución definitiva, ya que la política de prohibición, adoptada por Alemania, por ejemplo, es una solución a corto plazo; a medio y largo plazo, la tendencia es que la IA se expanda hacia el Derecho de manera inevitable. Ante esta posibilidad y la advertencia pesimista, desde una posición garantista, en la que se reconoce la necesidad de una respuesta del Derecho Público ante el cambio de paradigma, los algoritmos tendrían naturaleza jurídica de reglamentos, ya que, si se están utilizando como un código jurídico, pero en una versión avanzada, un código 2.0, como si fuera un reglamento administrativo, se deben aplicar las mismas garantías jurídicas ya existentes en el ordenamiento. La posición se justifica ante la comprensión de que tanto los reglamentos administrativos como los algoritmos son y serán utilizados para innovar en la formación del ordenamiento jurídico, ya sea para evaluar las circunstancias de casos concretos o para indicar cuál debe ser la respuesta jurídica adecuada a un caso dado presentado a la Administración Pública. Ante esta aplicación materialmente jurídica de los algoritmos, es necesario garantizar, con algunos cambios normativos: el procedimiento de creación, la publicidad y el control del código algorítmico, de la misma manera que se garantizan estos derechos para el reglamento tradicional.

El posicionamiento contrario a la idea de que los algoritmos sean reglamentos, o que tengan alguna naturaleza jurídica, parte de la concepción de que el concepto de algoritmos es demasiado amplio como para siempre tener la misma naturaleza o régimen jurídico aplicado a ellos. Según Huergo Lora (2020, p. 65), que los algoritmos sean reglamentos no limita su uso, como piensa Boix Palop (2020), sino que permite convertirlos en norma jurídica, sin que tengan forma necesaria para ello. Para evitar esta transmutación del código algorítmico en norma jurídica, la solución sería negar cualquier naturaleza jurídica a los algoritmos (HUERGO LORA, 2020: 66).

Berning Prieto (2023: 106) entiende que es necesario dotar el uso de algoritmos con garantías jurídicas. Tal propósito se lograría mediante la categorización de los algoritmos como lo que son: algoritmos. Para el autor, no es necesario clasificar los algoritmos dentro de algún instituto jurídico ya existente. La crítica a la idea de crear, dentro de la teoría jurídica, una nueva categoría para los algoritmos es la siguiente: la posición de Berning Prieto (2023) deja vulnerable el Derecho Público al afirmar que, en el actual ordenamiento jurídico, no existe la posibilidad de regular y garantizar derechos y garantías fundamentales sin la creación de un nuevo instituto, aplicado exclusivamente para la tecnología de IA. Además de desconsiderar los actuales mecanismos de protección al ciudadano, el camino propuesto deja la posibilidad de que, con cada nueva tecnología, el Derecho tenga que abrir espacio en su teoría para la creación de una nueva categoría, sin la cual la sociedad quedaría desprotegida.

Huergo Lora (2020) afirma que los algoritmos no tienen naturaleza jurídica; sirven para la mera instrumentalización de la administración pública o que son utilizados como administración

digital, ya que, en estos casos, al tratarse de mero instrumento o reflejo de la norma o acto administrativo (como ocurre con los programas de declaración de impuestos sobre la renta), al atacar la norma o acto administrativo, con el medio jurídico adecuado, se está atacando, también, por reflejo, al algoritmo. La verdadera discusión estaría en la adopción de algoritmos predictivos. La expresión "algoritmo predictivo", para Huergo Lora (2020), se le atribuye el significado de conocimiento, sin embargo, no se dice que se trata de tecnología. Además, hay un problema con la opción de la expresión "algoritmo" adjetivada con el término "predictivo": El término adoptado por Huergo Lora (2020) limita a solo un tipo de algoritmo, el predictivo, en el uso de la administración pública, cuando, en realidad, el uso de sistemas inteligentes por parte de la Administración Pública puede ser más amplio, incluyendo el algoritmo experto (HÄUSELMAN 2022). Ante las diversas posibilidades de utilización de la tecnología y los riesgos de estos usos, no reconocer la necesidad de establecer qué régimen jurídico aplicar al uso de IA en el Derecho Administrativo es abrir la posibilidad de aplicar diversos sistemas inteligentes, predictivos o no, en el cotidiano de la administración, sin establecer de manera uniforme y segura cuáles son las garantías y controles posibles que pueden ejercer los ciudadanos. Si no establecer una naturaleza jurídica para los algoritmos genera inseguridad, ¿dotarlos de naturaleza jurídica de regulación sería la solución más adecuada?

Según Boix Palop (2020), sí. Las críticas a establecer una naturaleza jurídica para la IA pueden ser de orden técnico y financiero, o de orden jurídico. Respecto a las críticas de orden técnico y financiero, como la indicación de que los lenguajes de la IA son incomprensibles para las masas, o que es costoso desde un punto de vista financiero implementar un sistema de participación, publicidad y control en la creación, implementación y uso de sistemas inteligentes, tenemos la respuesta de Boix Palop (2020: 265): tales críticas no niegan la visión pesimista de la adopción de la tecnología, ni cuestionan el argumento de que hay una naturaleza jurídica que debe ser encontrada y debatida; simplemente levantan barreras que intentan impedir la correcta aplicación del Derecho a la tecnología. Por lo tanto, consideramos, para fines de argumentación, que las dos variables, la técnica y la financiera, son posibles en la determinación de una naturaleza jurídica para la IA.

Para elaborar reglamentos, la Administración Pública cuenta con límites formales y materiales (GARCÍA DE ENTERRÍA; FERNÁNDEZ, 2017: 219 y ss.). Entre los límites formales se encuentran la competencia para emitir reglamentos, el respeto a la jerarquía normativa y la obediencia al procedimiento administrativo legalmente previsto, permitiendo incluso la participación de interesados (FERNÁNDEZ FARRERES, 2018: 572); y entre los límites materiales se incluyen el respeto a los actos y principios generales del derecho, el control de la discrecionalidad y el respeto a la materia que debe ser tratada mediante reglamentación. Al atribuir la naturaleza jurídica de reglamento a los algoritmos, se requieren cambios constitucionales y legislativos para evitar violaciones a los derechos y garantías fundamentales propios de los estados

modernos. Para que un algoritmo sea un reglamento, con validez normativa, no basta con cumplir el requisito material; también y necesariamente debe cumplir con el requisito formal, de lo contrario estaría violando el principio de legalidad, fundamento del Estado Democrático de Derecho. Un acto administrativo derivado de un reglamento hecho por un algoritmo sin cumplir las garantías legalmente previstas sería nulo.

Si el algoritmo se utiliza en un sistema experto, un instrumento para la Administración Pública, lo que se tiene es un uso del algoritmo para reproducir el reglamento, de manera matemática, con el fin de aumentar la eficiencia y eficacia de la Administración Pública en actividades plenamente regladas. En estos casos, aunque el algoritmo sea un reglamento, el algoritmo no creó el reglamento. García de Enterría y Fernández (2017, p. 209) recuerdan que el reglamento es toda norma escrita por la Administración Pública y el algoritmo no es la Administración Pública. Si el algoritmo se utiliza en un sistema que aprende, el Derecho Público tendría que renunciar a parte de su estructura tradicional para que la naturaleza jurídica de reglamento, atribuida a la IA, tenga legalidad y juridicidad: lo que sería un nuevo tiempo para el Derecho, ya que estaría transformándose, con el uso de la tecnología, en una nueva estructura, distinta de la existente desde los tiempos de la Francia posrevolucionaria (GARCÍA DE ENTERRÍA, 1994).

4. Anatomía del acto administrativo: la naturaleza jurídica de la IA para el Derecho Administrativo

En el actual sistema jurídico, para que exista la posibilidad de controlar los actos administrativos y garantizar los derechos, la IA, cuando es utilizada por la Administración Pública en el ejercicio de sus funciones, es el motivo del acto administrativo, aunque una parte de la IA, el algoritmo, actúe como reglamento del acto administrativo. La solución está dada; ahora argumentamos a favor de esta conclusión. Los países de tradición jurídica continental europea comparten una misma teoría general de Derecho Administrativo, que tiene su origen en la Francia posrevolucionaria (GARCÍA DE ENTERRÍA, 1994). Este conjunto normativo y organizacional, que establece una estructura fundamental y legal para el Derecho Administrativo (con teoría sobre las personas administrativas, sus poderes y bienes, la jerarquía, la teoría de los actos administrativos, reglamentos y responsabilidad), fue incorporado en el Derecho portugués y español, y también implementado en la Iberoamérica, donde se encuentran los actos administrativos.

Acto administrativo, en sentido amplio, es “todo acto jurídico dictado por la Administración y sometido al Derecho Administrativo” (GARCÍA DE ENTERRÍA; FERNANDEZ, 2017: 591), lo que lo distingue de la actuación judicial, de los actos practicados por los administrados y de los actos practicados mientras particulares. Según Bandeira de Mello (2015: 393), en traducción libre, el acto administrativo es la “declaración del Estado (...) en ejercicio de prerrogativas públicas, manifestada mediante providencias jurídicas complementarias de la ley para darle cumplimiento, y sujetas a control de legitimidad por órgano jurisdiccional”. Se trata de una

declaración jurídica del Estado, o de quien actúe en su nombre, que certifica, crea, extingue, transfiere, declara o modifica derechos y deberes. Esta declaración se ejerce en el uso de prerrogativas públicas y está regida por el Derecho Público, estando, en un Estado Democrático de Derecho, sujeta a control interno, por parte de la propia Administración Pública, y externo, por parte de órganos jurisdiccionales. El concepto amplio abarca tanto los actos concretos, que serían los actos administrativos en sentido estricto, destinados a aplicar el derecho objetivo, como los reglamentos, actos abstractos que pueden innovar en el derecho objetivo (GARCÍA DE ENTERRÍA; FERNANDEZ, 2017: 596).

En sentido estricto, en el cual se excluyen los contratos administrativos y los reglamentos, el acto administrativo sería “la declaración de voluntad, de juicio, de conocimiento o de deseo realizada por la Administración en ejercicio de una potestad administrativa distinta de la potestad reglamentaria” (GARCÍA DE ENTERRÍA; FERNANDEZ, 2017: 591). El acto es una actividad intelectual, producto de la inteligencia del sujeto competente, que actúa en nombre de una Administración Pública.

Un acto administrativo que ha pasado por todo el trámite de formación y que se ajusta a las exigencias normativas, produciendo sus efectos típicos, atípicos y reflejos, se considera un acto perfecto, válido y eficaz. Es dentro de esta idea de acto perfecto, válido y eficaz, donde se posiciona la IA como parte del acto. La teoría del acto administrativo se presenta en los manuales de forma más o menos completa, dependiendo del autor; por lo que no hay acuerdo, incluso dentro de un país, en cuanto a la terminología o cantidad de requisitos del acto. Presentaremos una anatomía del acto administrativo, analizando en partes sus requisitos, como abstracción, para identificar mejor la naturaleza jurídica de la IA para el Derecho Administrativo Iberoamericano.

Los requisitos del acto administrativo, sin los cuales no es posible hablar de legalidad de la actuación estatal, pueden dividirse en: elementos del acto administrativo, que son constitutivos de su existencia; y presupuestos del acto administrativo, que garantizan, en caso de ser un acto perfecto y válido, su permanencia en el ordenamiento jurídico. En cuanto a los elementos, son intrínsecos al acto: contenido y forma. Los requisitos que no son intrínsecos al acto son los presupuestos de existencia: objeto y pertinencia; y los presupuestos de validez: sujeto, causa, formalización, procedimiento, fin y motivo. Sin elementos no hay acto; sin presupuestos de existencia falta lo indispensable para la producción de efectos jurídicos; y sin los presupuestos de validez, no hay acto administrativo válido.

En cuanto a los elementos: el contenido es aquello que el acto decide, enuncia, certifica, crea, extingue, modifica u opina en el orden jurídico. Es el acto en su esencia. Si el contenido es ilícito, el acto es inválido y debe ser declarado nulo. La IA no puede ser el acto en su esencia, ni el enunciado, certificación, creación, extinción, modificación u opinión en el orden jurídico, ya

que en la organización administrativa existe la necesidad de un sujeto competente para pronunciar o suscribir el contenido del acto, razón por la cual, la IA no es el contenido o la materialidad del acto.

La forma, segundo y último elemento del acto administrativo, es el revestimiento exterior del acto, el modo en que la Administración Pública exterioriza el acto. Superada la fase de gestación del contenido del acto, es necesario dotarlo de una forma para que exista en el mundo. En el actual Derecho Administrativo, la forma se manifiesta en la formalidad del acto (presupuesto de validez), que, por razones de seguridad jurídica y publicidad, generalmente, es escrita; también se admite la forma verbal, imagética y/o sonora. La IA tampoco es la forma del acto, ya que la forma es simplemente la técnica de cómo se exterioriza la manifestación de voluntad.

En cuanto a los presupuestos de existencia: el objeto se refiere a aquello sobre lo que el acto dispone, debiendo ser material y jurídicamente posible. Según García de Enterría y Fernández (2017: 607), el objeto debe ser lícito, determinado y posible. Si no hay objeto en la realidad sobre el cual el acto vaya a producir efectos, como, por ejemplo, la destitución de un servidor ya fallecido, no hay objeto y, por lo tanto, no hay acto. La IA tampoco es el objeto del acto, ya que no es la realidad sobre la cual incide la manifestación unilateral de voluntad de la Administración Pública.

El segundo requisito de existencia es la pertenencia a la función administrativa, que establece que el acto administrativo es el ejercicio de una función administrativa. Es decir, solo la Administración Pública tiene competencia para ejercer esta función y llevar a cabo actos administrativos. Por lo tanto, la IA no constituye la pertenencia a la función, ni puede ser ejercida por particulares para la realización de actos administrativos.

A lo que toca a los presupuestos de validez, tenemos lo siguiente: el sujeto, que es quien realiza el acto administrativo. El sujeto debe ser competente para llevar a cabo el acto, además de no tener ningún impedimento legal para hacerlo, como, por ejemplo, estar de vacaciones, en licencia, o impedido por cualquier motivo legal. Dado que el sujeto es quien realiza el acto administrativo, no se confunde con el acto en sí mismo, su contenido o su forma, que son elementos del acto. La IA no es el sujeto del acto administrativo, ya que la Administración Pública, como institución jurídica, se manifiesta a través de sus órganos y agentes, en el ejercicio de sus competencias (por razón de materia, espacio o tiempo), previamente establecidas por ley. Ningún sistema inteligente tiene, en el actual ordenamiento jurídico, la competencia para ser el sujeto y llevar a cabo un acto sin la supervisión de un humano, como sujeto competente, al que se le imputará la responsabilidad.

El procedimiento, segundo presupuesto de validez, ocurre por imposición normativa o reglamentaria, siendo la manera de producir un acto (GARCÍA DE ENTERRÍA; FERNÁNDEZ, 1997: 608). Consiste en otros actos administrativos que deben ocurrir en un orden determinado,

hasta llegar a la resolución final, que permitirá la manifestación unilateral de la voluntad administrativa. El procedimiento no es un acto compuesto que une diferentes voluntades, sino un conjunto de actos, todos instrumentales o auxiliares del acto final. El procedimiento está relacionado con el sujeto, ya que los actos administrativos solo pueden ser realizados por el órgano competente. La IA no es el procedimiento, aunque puede formar parte de ello como una forma de producir conocimiento para ser aplicado a un fin determinado (el motivo del acto). El procedimiento es simplemente la técnica mediante la cual la Administración Pública elige manifestar su voluntad unilateral, mientras que la IA es una tecnología que produce conocimiento para ser utilizado en un procedimiento administrativo. El procedimiento debe ser llevado a cabo por un órgano competente, mientras que la IA produce una solución matemática que debe ser suscrita por el sujeto competente para realizar dicho acto.

El fin del acto es el bien jurídico objetivado por el mismo. El acto solo puede cumplir con el fin determinado por la norma. Por ejemplo, evacuar un área afectada por un evento climático grave tiene como fin salvar vidas. La explicación de este presupuesto de validez deja en claro la imposibilidad de que la IA sea el fin del acto administrativo, ya que la tecnología no es un fin en sí misma, sino el conocimiento que permite determinar, por ejemplo, qué área podría verse afectada por un evento climático, motivando así el acto administrativo de evacuación de la población.

La formalización es la forma específica en que el acto debe ser expresado. Por lo general, por razones de seguridad jurídica, el acto administrativo es escrito. Sin embargo, en ciertas ocasiones, puede ser verbal, visual y/o gráfico. La formalización también requerirá la presencia de otro requisito de validez: la motivación del acto. La IA no es la formalización del acto, ya que es una tecnología a ser aplicada por la Administración Pública y no simplemente una técnica para expresar unilateralmente la voluntad de la Administración Pública.

La causa es el requisito de validez que vincula el motivo del acto con el objeto del acto. Es decir, en traducción libre, "si el agente se basa en motivos que no guardan congruencia, pertinencia, con el acto que practicó, este estará viciado" (BANDEIRA DE MELLO, 2015: 416). Es en el presupuesto de validez de la causa donde se debe analizar la proporcionalidad del acto administrativo, de modo que los motivos sean ponderados en relación con el objeto. La IA no es causa, ya que no es ella quien pondera, sino que simplemente presenta un resultado. Quien pondera es el sujeto competente para la práctica del acto. Los actos administrativos, cuando están viciados en su causa, serán objeto de control a través del motivo del acto, siendo más una razón para no considerar a la IA como el motivo del acto. El último presupuesto de validez es el motivo, la naturaleza jurídica de la IA para el Derecho Administrativo.

El motivo es lo que autoriza o exige la práctica de determinado acto administrativo. Es el conocimiento aplicado a un fin que se debe tener en cuenta en la realización del acto

administrativo. Por ser externo y antecedente al propio acto, no puede ser un elemento, siendo, por tanto, un presupuesto. El motivo puede estar previsto en la ley, o no. Si hay un motivo legal, se tiene una previsión abstracta de una situación fáctica; mientras que el motivo del acto es la propia situación material que sirve de soporte real y objetivo para la realización del acto administrativo. Al no determinar la existencia del acto, como lo hacen el objeto, que existe en la realidad; y la prerrogativa de función pública, que determina el sujeto y la competencia, el motivo resulta ser un presupuesto de validez.

El motivo, como presupuesto de validez del acto administrativo, tiene sus propios elementos: el móvil, la motivación y la determinación del acto a partir del motivo presentado. La realidad objetiva y externa al sujeto competente para la práctica del acto administrativo (motivo del acto), tiene en el móvil del acto (intención del sujeto), la representación subjetiva, psicológica e interna del sujeto competente para la práctica del acto administrativo y corresponde a aquello que suscita la intención de dicho sujeto con la práctica del acto y lo lleva a ponderar, en la causa del acto, sobre su aplicación en la realidad. El móvil es relevante cuando hay discrecionalidad, pero está presente, también, en los actos vinculados.

La motivación integra la formalización del acto administrativo, como una exposición de los motivos, en la cual se enuncia la regla jurídica aplicada, los hechos en los cuales el sujeto se basó para decidir y la pertinencia lógica entre los hechos y el acto administrativo practicado. Navarro (2012: 905) define la motivación como “la exteriorización de las razones fácticas y jurídicas que conducen a la adopción de un determinado acto administrativo, con un determinado contenido”. La motivación es una imposición vigente desde el Derecho Público posrevolucionario, en Francia, siendo un componente central del sistema continental europeo de legalidad (GARCÍA DE ENTERRÍA, 1994: 172).

Según la teoría de los motivos determinantes, los hechos que sirvieron de soporte a la decisión integran la validez del acto, de modo que, si se comprueba la inexistencia o invalidez del motivo, se comprueba, también, la invalidez del acto. Es por medio del motivo que la Administración Pública ofrece “argumentación que deje constancia del proceso mental y jurídico que han conducido a toma de esa decisión, que permita a los destinatarios de lacto y a los órganos administrativos o judiciales que, en caso de recurso, deban enjuiciar su validez” (NAVARRO, 2023: 905). Motivar el acto es un medio técnico de control de la causa del acto (GARCÍA DE ENTERRÍA; FERNÁNDEZ, 2017: 614). Por esta razón, que permite un control efectivo del acto administrativo a partir de su motivo, la Administración Pública tiene el deber de motivar los actos administrativos. Al fin y al cabo, sin la motivación, se produce la nulidad absoluta de acto administrativo o resolución, por ausencia de un presupuesto de validez.

Cuando es utilizada por la Administración Pública, la IA, en cualquiera de sus sistemas inteligentes, proporciona al sujeto competente para la realización de un acto administrativo,

conocimientos que deben ser aplicados a un fin. La IA presenta a la Administración Pública una realidad, ya sea a través de una predicción o de una respuesta basada en aprendizaje, que servirá como motivo para la práctica del acto administrativo, vinculándose a la validez del propio acto y permitiendo el control de dicho acto, en caso de que los motivos presentados por un determinado sistema inteligente sean falsos, inverídicos, ilegales o, simplemente, no se adecuen a la causa del acto.

Las decisiones administrativas tradicionales se basan en normas y en la racionalidad simbólica del ser humano. Con la IA, las decisiones administrativas pasan a basarse en normas y en la racionalidad matemática del sistema inteligente; se añade un modelo matemático donde antes solo existía la intención del sujeto competente. Sin embargo, la opción de un sistema u otro no elimina la causalidad y la intención, solo ofrece una nueva forma de presentar los motivos del acto. Los sistemas inteligentes se fundamentan en datos, considerados por los programadores como indicadores (o proxies) de la situación a tratar. Estos indicadores serán presentados al administrador público, quien los utilizará como motivos del acto.

Ejemplos ayudarán a comprender por qué la IA, cuando es utilizada por la Administración Pública en el ejercicio de sus prerrogativas públicas, es el motivo del acto administrativo: En la Administración policial de los Estados Unidos, el sistema inteligente COMPAS, que trata de la reincidencia criminal (COECKELBERGH, 2023) y los sistemas inteligentes Predpol, HunchLab y Beware, que son predictivos de actividades delictivas (DÍAZ GONZÁLEZ, 2020: 185 y ss.) proporcionan conocimiento y datos a la Administración Pública, que, con este conocimiento, motiva el acto administrativo para actuar en la realidad. En España, para volver a la tradición continental de Derecho Administrativo, el sistema inteligente VERIPOL analiza el texto de las denuncias, orientando la intervención policial, con una tasa de acierto del 90% (DÍAZ GONZÁLEZ, 2020: 185). Nuevamente, la IA sirve como motivo del acto administrativo (intervención policial). Ben-Shahar (2021: 237), informa sobre la posibilidad de utilizar la IA para la concesión de créditos, de modo que el sistema inteligente analiza el perfil del ciudadano, informando al administrador público sobre las posibilidades de pago futuro. Una vez más, es el administrador público quien va a practicar el acto, estableciendo el motivo desde la IA para conceder, o no, el crédito. Los sistemas inteligentes de reconocimiento facial (HÄUSELMANN, 2022: 237) realizarán el reconocimiento de personas buscadas por el sistema judicial, pero es el administrador público, motivado por la información proporcionada por la IA, quien actuará, mediante acto administrativo, en la realidad, emitiendo la orden de arresto, búsqueda y captura o cualquier otra, requerida por el sistema jurídico. De la misma forma, el sistema inteligente VioGén, de España, proporciona motivos para la actuación administrativa con el fin de evitar la violencia de género. El sistema Alice informa al Tribunal de Cuentas de la Unión, en Brasil, sobre posibles irregularidades, para que el órgano de control pueda iniciar una investigación. Los sistemas Watson and Debater (ASHLEY, 2017: 28 y ss.) y de Neota Logic

(ASHLEY, 2017: 35), al presentar una respuesta jurídica a un problema propuesto, proporcionan al administrador público motivos para la práctica del acto administrativo.

4. Resolviendo el problema jurídico: La IA como el motivo del acto administrativo

Los diferentes sistemas inteligentes se dividen en dos grupos: los sistemas inteligentes que aprenden, utilizados en actos administrativos discrecionales, y los sistemas inteligentes expertos, sin capacidad de aprender, utilizados en actos administrativos vinculados. Ambos presentan el motivo del acto administrativo, pero, con vistas a garantizar los derechos de los ciudadanos, tienen mecanismos diferentes de implementación en la realidad en cuanto a control, publicidad y procedimiento, dentro del Derecho Administrativo de la tradición continental europea.

Si el sistema inteligente utilizado por la Administración Pública es un modelo de sistema experto, lo que se tiene, en esencia, es la traducción del reglamento en algoritmo (BOIX PALOP, 2020), transformando la IA en una técnica para la realización de las actividades administrativas. En estos casos, el algoritmo y el reglamento son idénticos, distinguiéndose solo por el lenguaje en el cual están contruidos. Pero fue la propia Administración Pública, por medio de su sujeto competente, presupuesto de validez del acto administrativo, en el ejercicio de su competencia, quien creó el reglamento, habiendo la IA solo hecho la reproducción. Esa reproducción es la que va a entregar, al agente competente, el motivo del acto administrativo.

El uso de sistemas expertos, como señala Carullo (2021), se produce en actos vinculados, tanto en sus presupuestos como en su contenido. El resultado esperado es el mismo en cada caso, sea la decisión de un ser humano o de una IA. Lo que resuelve la imputabilidad de la decisión automatizada: el sujeto competente está identificado desde el momento en que el órgano competente recibe el contenido de la decisión emitida por el sistema experto. El motivo del acto en sistemas expertos son los enunciados de hecho y de derecho que integran el caso, siendo irrelevante el instrumento utilizado para la correcta aplicación de la norma; lo importante es que la decisión esté de acuerdo con el reglamento y con la ley. La motivación de estos actos derivados de sistemas inteligentes expertos debe ser la misma, en contenido y forma, que la realizada por el sujeto competente, sin la ayuda de la IA.

En caso de litigio de la decisión administrativa realizada con la ayuda de un sistema experto, ya que se trata de un acto reglado, el sujeto competente puede repetir la toma de decisiones y verificar la validez y legalidad; momento en el cual la Administración Pública puede mantener el acto como perfecto, válido y eficaz, o ejercer el control, anulándolo. Si el acto se mantiene, por entendimiento de existencia de validez, al administrado aún le queda la revisión judicial, en la cual puede probar la ausencia o ilegalidad del motivo o de algún otro presupuesto de validez o existencia de lacto.

Los interesados también pueden solicitar acceso al algoritmo y al modelo, para verificar si es fiel, o no, al reglamento. En estos casos, la publicidad del algoritmo y modelo no necesita ocurrir junto con la motivación del acto. La publicidad puede darse solo si existe una impugnación a la decisión. Si se trata de código abierto, la solución es sencilla: basta dar acceso a quien lo solicite. Si se trata de código cerrado, en el cual es preciso respetar el derecho de propiedad sobre la IA, aun así, es posible dar acceso al algoritmo sin violar el secreto de empresa; ya que, con el recurso de programación denominado *differential privacy*, se altera ligeramente el algoritmo en el sistema inteligente, ofuscando datos sensibles, para acceso público, sin que sea posible para terceros la reproducción y explotación económica de la IA (ZHAO; CHEN, 2022).

Entre las dos soluciones, de código abierto o cerrado, para que se garantice la publicidad posterior a la ejecución del acto, si es posible tener una IA de código abierto, la Administración Pública puede exigir tal característica en el pliego de licitación o en la contratación directa. Si existe estructura económica y tecnológica, la propia Administración Pública puede construir su código, de forma abierta, para publicarlo, en casos de litigio. Si no es posible el código abierto, aun así, la publicidad puede ser garantizada, con el acceso a partes del algoritmo (*differential privacy*), en especial a las partes relativas al motivo, si la Administración Pública dispone de medios para ello, o se está obligada a cumplir con una orden judicial para la divulgación del algoritmo o de partes de ello.

La publicidad puede facilitar ataques al sistema inteligente, que, a pesar de los niveles de seguridad, puede tener limitaciones en cuanto a la programación, como ocurre, en especial, con los sistemas de red neuronal (MITCHEL, 2019: 102). Szegedy et al. (2014) explican que, en una red neuronal de lectura de imagen, con un simple cambio de píxeles, imperceptible para el ojo humano, es posible alterar la respuesta presentada por el sistema inteligente, generando falsos positivos o falsos negativos. En sistemas inteligentes que van a leer documentos, por ejemplo, para que la Administración Pública conceda o niegue un beneficio, la integridad del sistema es importante, para evitar fraudes, de modo que la parcialidad de la publicidad sería justificable. En estos casos, la Administración Pública puede considerar qué partes del algoritmo o modelo se publicarán, para garantizar, al mismo tiempo, el cumplimiento de la publicidad y la integridad del sistema.

La lectura del código, en un contencioso administrativo o judicial, es una prueba técnica, de modo que, además del código, el responsable del sistema inteligente (particular o Administración Pública), debe ofrecer las herramientas e instrucciones necesarias para la comprensión del lenguaje computacional, en especial sobre cómo se llegó al motivo. Esta exigencia responde a la idea de que la publicidad debe ser accesible, aunque técnicamente no todos puedan comprender el contenido de los documentos presentados.

Ahora, el segundo grupo: los sistemas inteligentes que aprenden. Estos sistemas de aprendizaje adoptan un sistema no procedimental para llegar a una respuesta, lo que los hace no verificables por un ser humano, fenómeno al que se le llama *black box* (CARULLO, 2021: 444). Estos sistemas inteligentes se entrenan con datos, creando un modelo que será utilizado en los casos concretos. Los datos para el entrenamiento deben estar anonimizados y bajo el control de la Administración Pública. Una vez entrenados, los sistemas inteligentes pueden, por sí solos, presentar una respuesta, de acuerdo con el modelo matemático. Como las reglas (modelo matemático) se crean durante el proceso de aprendizaje del sistema inteligente, es la propia máquina, y no el sujeto competente, quien crea los parámetros con los cuales la Administración Pública tendrá que motivar el acto. Es como si el sistema inteligente, una vez entrenado, empezara a decidir solo, en un sistema de precedentes administrativos, creando un modelo capaz de ofrecer a la Administración Pública un escenario para evaluación y aplicación en casos futuros y semejantes. Sin embargo, en el actual ordenamiento jurídico, de tradición continental europea, el sistema inteligente no puede decidir solo: de acuerdo con la legalidad, la imputabilidad del acto será del sujeto competente que suscribirá el acto.

Para que haya control del acto administrativo realizado con la asistencia de un sistema inteligente de aprendizaje, mientras sea motivo del acto administrativo, la respuesta deberá presentar, además de la respuesta binaria de sí/no para el caso concreto, un texto lógico-jurídico, de forma articulada, tal como lo haría un ser humano, con la misma consistencia de los datos con los que fue entrenado, para que la Administración Pública pueda motivar el acto resultante del uso del sistema inteligente. Sin el texto, al acto administrativo le faltaría la motivación y, por lo tanto, uno de sus presupuestos de validez. La posibilidad de que el sistema inteligente ofrezca el modelo interpretable debe constar en el pliego de licitación, o en el contrato, en caso de ser necesario contratar con particulares; si la IA es creada por la propia Administración Pública, con el objetivo de dar publicidad posterior a la motivación del acto administrativo, debe constar la posibilidad de exhibición de un modelo interpretable en las directrices de construcción del sistema inteligente.

Con sistemas inteligentes de redes neuronales, la tendencia es que haya mucha complejidad, de modo que sea imposible la entrega de un modelo interpretable (SOUZA, 2015). En estos casos, la Administración Pública puede utilizar la respuesta proporcionada por el sistema inteligente como un informe que motiva el acto, manteniéndose dentro de la legalidad prevista en el actual arreglo del Derecho Público. Esta consulta hecha a la IA, como tradicionalmente se consulta a órganos técnicos que responden con informes, aunque indique motivos para la práctica de determinado acto administrativo, no tendría fuerza vinculante y podría ser litigada, con la mera comprobación de que las razones de hecho y de derecho expuestas por la IA no cumplen con lo presupuesto de validez del acto administrativo. Otra solución, más extrema y que considera las visiones pesimistas del uso IA, sería seguir el ejemplo alemán (HUERGO LORA,

2020: 79) y prohibir el uso de sistemas inteligentes que no permiten la auditoría de cómo se tomó la decisión.

Además del texto articulado, para la validez del motivo, la Administración Pública debe establecer un nivel de confiabilidad de la respuesta, sin el cual el acto sería practicado en exceso de poder, violando, una vez más, la validez del acto administrativo. Aunque el grado de confiabilidad de la respuesta esté de acuerdo con lo establecido por la Administración Pública para la validez del acto, aún es posible ejercer el control sobre la relevancia de los datos utilizados en la construcción de la respuesta presentada por el sistema inteligente. Dado que se trata de un número elevado de datos, se puede exigir al sistema inteligente la presentación de los casos más relevantes utilizados en la construcción de esa respuesta. Con los casos más relevantes para la respuesta en mano, el sujeto competente puede evaluar la respuesta para evitar que el motivo presentado por los precedentes no tenga relación con el caso concreto en análisis. Si los datos precedentes no tienen relación con el caso concreto en análisis, faltaría el motivo del acto administrativo, sin el cual faltaría también un presupuesto de validez.

El hecho de garantizar la publicidad no significa, en abstracto, que la publicidad del sistema inteligente sea absoluta; debe someterse a la proporcionalidad y no ser utilizada si viola otros derechos y garantías fundamentales o si pone en riesgo la seguridad del propio sistema inteligente. En cuanto a la protección de otros derechos y garantías fundamentales, en España, el sistema VioGén, utilizado en la lucha contra la violencia de género, por ejemplo, no tiene sus datos, algoritmo y modelo divulgados por el Ministerio del Interior debido a que el sistema inteligente gestiona datos que tratan de la privacidad e intimidad de los ciudadanos. Tal expediente cumple con lo dispuesto en el Reglamento (UE) 2016/679 del Parlamento Europeo, en cuanto a la protección y gestión de datos. Norma que debe ser reflejada en toda Iberoamérica, cuando se aborde el tema, como hizo Brasil con la Ley n.º 13.709 de 2018.

Si los datos necesitan ser gestionados por más de una administración pública, como por ejemplo en un convenio administrativo entre municipios, para garantizar la protección de los datos de los ciudadanos, el sistema inteligente de federate learning (MAMMEN, 2024) puede ser adoptado para asegurar que solo el ayuntamiento en el que reside ese ciudadano tenga acceso a los datos. En términos de control y publicidad, por lo tanto, lo ideal es que el sistema inteligente proporcione indicadores del nivel de confianza de la respuesta, acceso al código fuente y al conjunto de datos – debidamente anonimizados –, utilizados en el entrenamiento del sistema.

Todas estas exigencias pueden figurar en el pliego de licitación, en caso de que haya contratación con particulares u otra administración pública; o formar parte de las directrices internas de la Administración Pública, en la construcción de sus propias inteligencias artificiales. Al igual que ocurre con los sistemas expertos, a efectos de publicidad, basta con el acceso a

partes del algoritmo, modelo y datos (ya que sistemas muy grandes tendrán millones – o más – , de datos). El acceso a parte del algoritmo o informes sobre el resultado presentado se justifica, ya que, cuando son utilizados por la Administración Pública, pasan a formar parte del contenido volitivo del motivo del acto administrativo. Si es imposible conocer estos elementos, se deja de cumplir con el motivo, requisito de validez del acto, lo que vuelve inválido el acto administrativo.

Más allá de todos estos mecanismos jurídicos, ya existentes en los ordenamientos de la tradición continental europea, el uso de una IA no impide que el ciudadano, que ha visto su derecho denegado en un acto administrativo realizado de forma automatizada, solicite, en juicio, la anulación del acto. Lo que se exige es la comprobación de hecho y de derecho de que la Administración Pública decidió mal o contra la ley, independientemente del uso de la tecnología. La anulación judicial del acto no depende del acceso al algoritmo, código fuente, modelo o datos anonimizados. La anulación del acto realizado con ayuda de la IA debe ocurrir con la comprobación de la inexistencia, inadecuación o ilegalidad del motivo presentado, o de otros requisitos del acto administrativo, como, por ejemplo, la ausencia de un sujeto competente en la práctica del acto, tal como se ha hecho durante mucho tiempo en actos realizados sin la ayuda de la IA.

5. Conclusión

La Revolución Francesa inauguró una época en la historia humana que aún está en expansión, tanto geográfica como fundamentalmente. Entre los logros de la revolución se encuentra la creación de una estructura administrativa organizada, articulada, eficiente, prestadora de servicios públicos y creadora de un espacio seguro para el ejercicio de la libertad. Para que este entorno administrativo prosperara, fue necesario desarrollar también una estructura de protección para los ciudadanos y de control de los actos administrativos. Esta estructura se ha adaptado y ha sobrevivido al sistema productivo y sus revoluciones industriales sin grandes alteraciones en sus mecanismos de funcionamiento.

La adopción de la IA, como tecnología, con contenido ético y político, capaz de alterar la forma en que la sociedad, el mercado y la Administración Pública se relacionan, presenta razones suficientes para justificar la adopción de un sistema de control y garantía de derechos. Después de todo, la IA es una tecnología capaz de alterar la realidad en la que vivimos, como lo fue, en los mitos, el fuego de Prometeo. En la mitología griega no existían mecanismos de control: la humanidad, al obtener el fuego, se volvió poderosa frente a los demás animales y frente a la naturaleza. La solución divina fue el castigo: encadenar a Prometeo y enviar la caja, con Pandora, para provocar males a la humanidad. Nosotros, en cambio, sí contamos con medios de control, para evitar que esta tecnología nos cause los males que los mitos anticipan. La implementación de este mecanismo de control, en el Derecho Público, pasa por la comprensión

de qué problemas jurídicos puede causar la adopción de la IA en su uso por parte de la Administración Pública; en especial, en cuanto a la posibilidad de control de los actos administrativos.

Las técnicas jurídicas de control de los actos administrativos, aunque admiten margen de mejora, ya existen, y actúan como un Hércules que libere a Prometeo de su castigo divino, pues se trata de utilizar esa tecnología de acuerdo con los derechos y garantías fundamentales garantizados por el Estado Democrático de Derecho. El sistema de control y garantía de derechos puede aplicarse independientemente de la tecnología adoptada por la Administración Pública, ya que el Derecho, como sistema tecnológico, juridifica la tecnología de la IA dentro de sus términos. La adopción de la IA no cambia la forma en que la sociedad ejerce control sobre sus actos: si antes, para la toma de decisiones, se podía recurrir a estudios científicos, estadísticas, informes de órganos consultivos y expertos, ahora, se puede recurrir a la IA. Se cambian los medios para obtener el motivo del acto, pero no se alteran los mecanismos de control jurídico para garantizar derechos y garantías fundamentales.

Bibliografía

ASHLEY, Kevin D. Artificial Intelligence and legal analytics: new tools for law practice in digital age. Cambridge: Cambridge University Press, 2017.

BANDEIRA DE MELLO, Celso Antônio. Curso de Direito Administrativo. 32. ed. São Paulo: Malheiros, 2015.

BEN-SHAHAR, Omri; PORAT, Ariel. Personalized Law: diferente rules for diferente peolpe. New York: Oxford University Press, 2021.

BERNING PRIETO, Antonio David. La naturaliza jurídica de los algoritmos. In GAMERO CASADO, Eduardo; PÉREZ GUERRERO, Francisco L. (coord.). Inteligencia artificial y sector público: retos, limites y médios. Tirant lo Blanch: València, 2023.

BOIX PALOP, Andrés. Los algoritmos son reglamentos. Revista de Derecho Publico: Teoria y Método. Vol. 1. 2020, pp. 223-270. Madrid, 2020.

CARULLO, Gherardo. Decisione Amministrativa e intelligenza artificiale. Diritto dell'informazione e dell'informativa. Fasc., 3, p. 432-461, 2021.

COECKELBERGH, Mark. La filosofia política de la inteligência artificial: uma introducción. Cátera: Teorema, 2023.

CUSTERS, Bart; FOSCH-VILLARONGA, Eduard (editors). Law nad artificial intelligence: regulating AI and applying AI in Legal Pratice. Berlin: Springer, 2022.

DÍAZ GONZÁLEZ, Gustavo Manuel. Algoritmos y actuación policial: la policía predictiva. In HUERGO LORA, Alejandro; DÍAZ GONZÁLEZ, Gustavo Manuel (coords.). La regulación de los algoritmos. Thompson Reuters: Navarra, 2020.

FALK, Svenja; RÖMMELE, Andrea; SILVERMAN, Michael. The Promisse of Digital Government in FALK, Svenja; Römmele, Andrea; SILVERMAN, Michael (editors). Digital Government: Leveraging Innovation to Improve Public Sector Performance and Outcomes for Citizens. Berna: Springer, 2017.

FERNANDEZ FARRERES, Germán. Sistema de Derecho Administrativo I. 4º ed. Thomson Reuters: Pamplona, 2018.

GARCÍA DE ENTERRÍA, Eduardo. La lengua de los derechos: la formación del Derecho Público europeo tras la Revolución Francesa. Alianza Editorial: Madrid, 1994.

GARCÍA DE ENTERRÍA, Eduardo; FERNÁNDEZ, Tomás-Ramon. Curso de Derecho Administrativo I. Thomson Reuters: Pamplona, 2017.

HÄUSELMAN, Andreas. Disciplines of AI: an overview of approaches and techniques. In CUSTERS, Bart; FOSCH-VILLARONGA, Eduard (editors). Law and artificial intelligence: regulating AI and applying AI in Legal Practice. Berlin: Springer, 2022.

HUERGO LORA, Alejandro. Una aproximación a los algoritmos desde el derecho administrativo. LORA, Alejandro Huergo; GONZÁLEZ, Gustavo Manuel Díaz (coords.). La regulación de los algoritmos. Thompson Reuters: Navarra, 2020.

KLEIJN, Roy de. Artificial Intelligence versus biological intelligence: a historic overview. In CUSTERS, Bart; FOSCH-VILLARONGA, Eduard (editors). Law and artificial intelligence: regulating AI and applying AI in Legal Practice. Berlin: Springer, 2022.

LELIS, Davi Augusto Santana de. Nem rápido, nem devagar: a natureza jurídica da inteligência artificial no Direito Administrativo europeu continental. *Revista de Informação Legislativa: RIL*, Brasília, DF, v. 62, n. 247, p. 37-63, jul./set. 2025. DOI: https://doi.org/10.70015/ril_v62_n247_p37. Disponível em: https://www12.senado.leg.br/ril/edicoes/62/247/ril_v62_n247_p37

MAMMEN, Pryanka Mary. Federate learning: opportunities and challenges. Disponível em: <https://arxiv.org/abs/2101.05428>. Acesso: 20-05-2024.

MITCHEL, Melanie. Artificial intelligence: a guide for thinking humans. Farrar, Straus and Giroux: New York, 2019.

NAVARRO, Karlos. Derecho Administrativo nicaragüense. In VARAS IBÁÑEZ, Santiago González. Derecho Administrativo em Iberoamérica (2ª ed.). Instituto Nacional de Administración Pública: Madrid, 2012.

PAIVA, José Eustáquio Machado. Um estudo acerca do conceito de tecnologia. Educação e Tecnologia. Belo Horizonte, jan/dez. 1999.

SORIANO ARNANZ, Alba. Creating non-discriminatory artificial intelligence systems: balancing the tensions between code granularity and general nature of legal rules. *Revista D'Internet, Dret i Política*, n. 38, 2023.

SOUZA, Paulo Vitor de Campos. Avaliação de técnicas de regularização aplicadas as redes neurais nebulosas (dissertação). LEMOS, André Paim (orientador). Programa de pós-graduação em engenharia elétrica: Universidade Federal de Minas Gerais: Belo Horizonte, 2015.

SZEGEDY, Christian; ZAREMBA, Wojciech; SUTSKEVER, Ilya; BRUNA, Joana; ERHAN, Dumitru; GOODFELLOW, Ian; FERGUS, Rob. Intriguing Properties of neural networks in Proceedings of the international conference on learning representations. 2014.

VARAS IBÁÑEZ, Santiago González. Derecho Administrativo en Iberoamérica (2ª ed.). Instituto Nacional de Administración Pública: Madrid, 2012.

VERASZTO, Estéfano Vizconde; SILVA, Dirceu; MIRANDA, Nonato Assis. Tecnologia: buscando uma definição para o conceito. Prisma. n.8, 2009.

ZHAO, Ying; CHEN, Jinjun. A survey on differential privacy for unstructured data content. ACM Computing Surveys, Vol. 54, n. 10s, Article 207. 2022.

Paula María Serrano Casaus. Protección de los Derechos Humanos de las personas extranjeras en los procedimientos administrativos de extranjería: Actuaciones automatizadas

Paula María Serrano Casaus

Universitat Jaume I

[ACCEDE A LA GRABACIÓN](#)

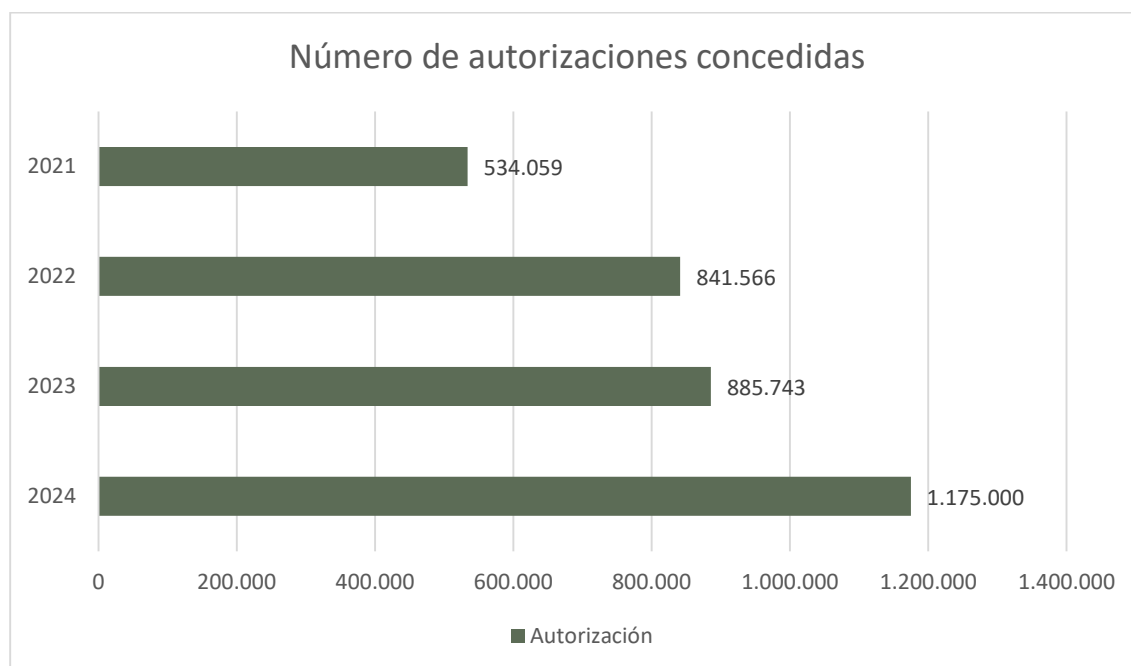
La inmigración en España es una realidad en constante evolución, impulsada por motivos laborales, familiares y humanitarios. Esta situación junto con la actual reforma del Reglamento de extranjería aprobado por Real Decreto 1155/2024, de 19 de noviembre, por el que se aprueba el Reglamento de la Ley Orgánica 4/2000, de 11 de enero, sobre derechos y libertades de los extranjeros en España y su integración social, ha influido en que se hayan introducido nuevas tecnologías en la Administración Pública de extranjería debido al aumento de solicitudes tramitadas desde España para que las personas extranjeras puedan adquirir la residencia temporal o la residencia y trabajo.

El uso de estas nuevas tecnologías puede contribuir a resolver los procedimientos en plazo, dada la gestión elevada que tienen las Subdelegaciones del Gobierno a través de las Oficinas de Extranjería, para resolver las solicitudes de autorizaciones administrativas de personas extranjeras.

Los sistemas de inteligencia artificial en la gran amplitud y diversidad de aplicaciones y funcionalidades que abarca pueden ofrecer a las Administraciones Públicas oportunidades para un cumplimiento más objetivo, eficaz y eficiente de los intereses generales. Siempre y cuando ello no suponga un riesgo para el pleno cumplimiento del principio de legalidad y para los derechos de las personas extranjeras, fundamentalmente en la esfera de los derechos fundamentales y Derechos Humanos. Deben ser sistemas transparentes dada la posible vulneración que podría darse a tales derechos si la Administración toma una decisión que no debería haber realizado con aplicación de actuaciones administrativas automatizadas, tal y como permite la Ley 40/2015, de 1 de octubre, de Régimen Jurídico del Sector Público.

Régimen jurídico aplicable a los extranjeros en España

Según los datos del Observatorio Permanente de la Inmigración del Ministerio de Inclusión, Seguridad Social y Migraciones, en el año 2024 se concedieron 1.466.639 autorizaciones de residencia a personas extranjeras, de entre las cuales, 1.175.000 corresponden a autorizaciones de residencia en régimen de extranjería, como se observa en el cuadro que se muestra a continuación²²⁰. Estas cifras denotan el aumento de solicitudes de autorizaciones de extranjería tramitadas a través del correspondiente procedimiento administrativo.



Si bien, es menester diferenciar entre término inmigrante y el término persona extranjera. El término inmigrante se engloba desde un contexto social, como aquella persona que abandona su país de origen para establecerse en otro país distinto. En cambio, persona extranjera, se puede conceptualizar desde un contexto jurídico, esto es, como aquella persona que tiene derechos y obligaciones, en función del territorio y atendiendo a la siguiente clasificación entre: personas que son nacionales de países de la Unión Europea; aquellas nacionales de países no integrantes de la Unión Europea (personas extracomunitarias) y los residentes de larga

²²⁰ Observatorio Permanente de la Inmigración de la Secretaría de Estado de Migraciones. *Flujo de autorizaciones de residencia concedidas a extranjeros. Año 2024. [en línea]. Septiembre 2025. [Consulta: 18 de septiembre de 2025]. Disponible en: https://www.inclusion.gob.es/web/opi/estadisticas/catalogo/flujo_autorizaciones*

duración. Estos últimos son aquellas personas extracomunitarias que han residido de forma legal en la UE durante más de cinco años²²¹.

Tal clasificación se hace necesaria en aras de determinar la referencia a personas extranjeras en España haciendo alusión a las personas nacionales de terceros países, por ser a aquellos a los que les es de aplicación la normativa de extranjería tal como la actual Ley Orgánica 4/2000, de 11 de enero, sobre derechos y libertades de los extranjeros en España y su integración social (en adelante Ley de extranjería) y el también vigente Real Decreto 1155/2024, de 19 de noviembre, por el que se aprueba el Real Decreto 1155/2024, de 19 de noviembre, por el que se aprueba el Reglamento de la Ley de extranjería (en adelante Reglamento de extranjería), dado que a los nacionales de países de la UE no les es de aplicación el citado régimen normativo, ni a los nacionales de los Estados con quienes los Estados miembros de la Unión Europea mantengan relaciones en virtud de Acuerdos Internacionales, siéndoles esta Ley de aplicación en los aspectos que pudieran ser más favorables tal y como dispone el artículo 1.3 LOEX²²².

La Ley de extranjería es ley especial según determina la propia Ley 39/2015, de 1 de octubre, del procedimiento administrativo común de las Administraciones Públicas (en adelante LPAC), tal y como lo especifica su Disposición adicional primera, al regular las especialidades por razón de la materia, concretamente “los procedimientos administrativos regulados en leyes especiales por razón de la materia que no exijan de los trámites previstos en esta ley o regulen trámites adicionales o distintos se regirán, respecto a estos, por lo dispuesto en dichas leyes especiales”. Esto significa que la Ley del procedimiento administrativo común se aplicará para aquello no regulado en la Ley de extranjería, o lo que es lo mismo, la ley administrativa es de carácter supletorio.

Las situaciones administrativas de los extranjeros en España

Según la normativa de extranjería, las personas extranjeras deben solicitar la correspondiente autorización para regularizar su situación administrativa, mediante solicitud en el Registro de los órganos administrativos competentes para tramitarlas, como las Delegaciones y Subdelegaciones del Gobierno, a través de las Oficinas de extranjería, a pesar de que la propia Ley de extranjería establezca entre sus objetivos el dotar a los extranjeros de un amplio abanico

²²¹ Clasificación efectuada por el Instituto Nacional de Estadística. [en línea]. Septiembre 2025. [Consulta: 18 de septiembre de 2025]. Disponible en:

https://ine.es/ss/Satellite?param1=PYSDetalleFichaSeccionUA&c=INESeccion_C&param3=1259944622628&p=1254735110672&pagename=ProductosYServicios%2FPYSLayou&cid=1259951667911&L=0

²²² ESPAÑA. Ley Orgánica 4/2000, de 11 de enero, sobre derechos y libertades de los extranjeros en España y su integración social. Boletín Oficial del Estado, 12 de enero del 2000, número 10, 54 páginas.

de derechos y libertades, tal como ejercer los derechos reconocidos en la Ley en las mismas condiciones que los españoles. De esta forma, tendrán una serie de derechos en función a su estatus jurídico, como puede ser, inmigrante regular, irregular, reagrupado, inmigrante irregular empadronado o inmigrante irregular no empadronado (Añón 2005).

Por ello al hablar de las situaciones jurídico-administrativas de los extranjeros en España se refiere al tipo de autorización que tienen para poder residir en España, o residir y trabajar, incluso estudiar, clasificando tales situaciones atendiendo a su tipo de residencia regular o irregular en España.

Los procedimientos de extranjería digitalizados.

Los procedimientos de extranjería se han ido acomodando a la realidad social actual. La implantación de la digitalización en el sector público es acopio de ello y propiamente en materia de extranjería, las propias Delegaciones y Subdelegaciones del Gobierno a través de las Oficinas de extranjería hacen uso de las herramientas digitales necesarias para tramitar y resolver los procedimientos administrativos de extranjería.

La comparecencia personal del interesado o de aquella persona legitimada para iniciar un procedimiento deja de ser necesaria al implantarse la tramitación de las solicitudes de las autorizaciones de extranjería de forma telemática a través de entre otros medios telemáticos, del Portal de extranjería del Ministerio de Política Territorial, tal y como dispone la Disposición adicional cuarta del Reglamento de extranjería²²³.

A través de la plataforma del Ministerio de Política Territorial, conocida como plataforma Mercurio, pueden realizarse diferentes solicitudes de forma telemática como pedir cita previa ante extranjería, consultar el estado en que se encuentra el expediente, pagar las tasas correspondientes a las autorizaciones, presentar las solicitudes de las autorizaciones de extranjería o renovar éstas, aportar documentos a expedientes que se están tramitando y presentar colectivamente la gestión de la contratación en origen.

La digitalización de los procedimientos y la automatización agiliza los trámites en materia de extranjería. Ejemplo de ello se encuentra en las concesiones de la nacionalidad española por

²²³ Disposición Adicional Cuarta del Reglamento de extranjería dispone “[...] Se desarrollará un sistema de información que incluya una aplicación informática común que permitirá la tramitación de todos los procedimientos de extranjería regulados en este reglamento y que registre la información y datos relativos a las personas extranjeras y ciudadanos de la Unión Europea residentes en España, sus autorizaciones y visados de larga duración”.

residencia en las que los interesados han visto reducido considerablemente el tiempo de resolución.

La incidencia de los procedimientos automatizados en los Derechos Humanos de las personas extranjeras

Los interesados pueden acceder a las Administraciones de extranjería a través de herramientas y plataformas digitales por las que se tramitan los procedimientos de extranjería. El acceso a estos procedimientos o servicios ofrecidos de forma telemática conlleva que las personas no tengan que acudir de forma presencial a las oficinas de extranjería.

La digitalización de las Administraciones Públicas a través de sistemas de Inteligencia Artificial (IA) ha generado una cierta agilidad en trámites de extranjería, pero con los riesgos que puedan generarse. El plazo en que deben resolverse las solicitudes de las autorizaciones de extranjería está regulado en el Reglamento de extranjería, si bien, es una realidad que las Oficinas de Extranjería, resuelven los expedientes tras haber finalizado el plazo concedido para ello. Es por este motivo, que la automatización de actuaciones sirve para agilizar los procedimientos y poder así ajustar la práctica administrativa con los principios generales que rigen la actuación administrativa como el de eficacia y eficiencia recogidos en el artículo 3 de la Ley 40/2015, reguladora del Régimen Jurídico del Sector Público. La agilización de los expedientes trae causa en el uso de una tecnología que automatiza procesos robóticos, como la revisión de formularios y la generación de respuestas estándar, o incluso generar a través de los datos personales aportados notificaciones de requerimiento de documentación en los expedientes de extranjería, como pueda ser la presentación de la tasa correspondiente que deba aportarse al expediente. La IA procesa grandes cantidades de información, y por ello es más eficiente en la prestación de servicios, reduciendo, además, los posibles errores que puede cometer el empleado público, por ser estos más rigurosos y precisos. Además, también revisa y clasifica información. Incluso identifica patrones y anomalías en los documentos, lo que también ayuda a detectar posibles fraudes o errores.

En primer término, no existe una definición unánime del término Inteligencia Artificial, pero podemos entenderla como “un conjunto de tecnologías que utilizan datos combinados y aplican modelos o algoritmos, junto con el poder de cómputo proporcionado por las Tecnologías de la Información (TI), para procesar y generar información de manera eficiente a partir de gran cantidad de datos de entrada o salida”. (La Espina 2024).

Por el contrario, si existe un término de actuación administrativa automatizada recogido en el artículo 41 de la Ley 40/2015, por el que “se entiende por actuación administrativa automatizada, cualquier acto o actuación realizada íntegramente a través de medios electrónicos por una Administración Pública en el marco de un procedimiento administrativo y en la que no haya intervenido de forma directa un empleado público”.

O también puede definirse desde la perspectiva de su naturaleza como “la actividad llevada a cabo por una Administración pública en el ejercicio de una potestad administrativa en función de las competencias que tiene encomendadas. La diferencia respecto a cualquier otra actividad realizada por medios convencionales radica precisamente en su fuente de producción. En este caso, la producción corresponde a un sistema de información programado y a no a una persona física. Ahora bien, la ausencia de persona física no es total, sino en cada caso singular” (Martín 2009).

De tal definición, podemos entender que son tres los requisitos que han de darse para que podamos entender que existe actuación administrativa automatizada: a) la actuación ha debido producirse a través de medios electrónicos, b) en el marco de un procedimiento administrativo y c) sin la intervención directa de un empleado público.

Sin embargo, en la esfera de los derechos fundamentales y los derechos humanos de todas las personas extranjeras cuando solicitan autorizaciones de extranjería, se debe prestar especial atención a que el uso de sistemas de inteligencia artificial no vulnere tales derechos. Especialmente, porque todavía presenta un desafío el tema de seguridad, transparencia y protección de datos personales en la IA. El Reglamento de la IA de la Unión Europea enumera una serie de normas cuya finalidad es el respeto y protección de estos derechos, prohibiendo el uso de toda aquella actuación automatizada que genere riesgos que afecten a tales derechos.

El derecho a la no discriminación por razón de género o raza consagrado entre otra normativa, en la Carta de los Derechos Fundamentales de la Unión Europea²²⁴, es especialmente relevante en los procedimientos de extranjería, si tenemos en consideración si una decisión que afecte a la situación jurídico-administrativa de una persona puede o no ser tomada a través de una actuación automatizada a través del uso de algoritmos generando sesgos que puedan dar lugar a discriminaciones. (Cernillo 2024).

En la misma línea se regula el derecho a la igualdad y a la no discriminación en el entorno digital que consagra la Carta de Derechos Digitales, al considerar que “el derecho y el principio a la igualdad inherente a las personas será aplicable en los entornos digitales, incluyendo la no discriminación y la no exclusión”. Así como establecer medidas específicas para evitar sesgos de género en los datos y en los algoritmos. Carta de Derechos digitales (España, 2021). En la Carta de Derechos Digitales se establece la obligación de evaluar el posible impacto a los derechos

²²⁴ El artículo 21.1 de la Carta de Derechos Fundamentales de la Unión Europea dispone textualmente “Se prohíbe toda discriminación, y en particular la ejercida por razón de sexo, raza, color, orígenes étnicos, o sociales, características genéticas, lengua, religión o convicciones, opiniones políticas o de cualquier otro tipo, pertenencia a una minoría nacional, patrimonio, nacimiento, discapacidad, edad u orientación sexual”.

digitales cuando se usan algoritmos en la toma de decisiones automatizadas o semiautomatizadas. O el propio derecho fundamental que también recoge la citada carta a la protección de datos de carácter personal por el mero hecho de ser persona.

En este contexto de protección de los derechos humanos y fundamentales a través de la aplicación de sistemas de Inteligencia Artificial, con la finalidad de evitar abusos y garantizar que su desarrollo y uso respeten la dignidad humana, el Parlamento Europeo dictó en 2017 una propuesta de carta sobre robótica, la cual contenía una serie de principios éticos por la que debía regirse la IA, entre otros, la justicia, el respeto de los derechos fundamentales, la transparencia o la rendición de cuentas. (Taboada 2024). Este último principio como una forma de protección de tales derechos.

Finalmente, la Comisión Europea en 2021 dicta la Ley de Inteligencia Artificial, pretendiendo regular los sistemas de IA más allá de los principios éticos, sino estableciendo obligaciones garantes de los derechos humanos y del uso seguro de la IA.

La Ley de Inteligencia Artificial no hace sino regular y clasificar los sistemas de IA en base a los riesgos que pueden generar en los derechos humanos.

Se ha de poner el foco en el ámbito de extranjería por ser los interesados en el procedimiento personas vulnerables sometidas a mayores discriminaciones por el mero hecho de ser inmigrantes. De esta forma, la gran cantidad de datos personales que tienen de ellos las administraciones de extranjería hace que sean altos los riesgos a los que están sometidos sus derechos humanos y así lo determina la cita ley.

Siguiendo con el respeto a los derechos humanos, en el año 2023, el Consejo de Derechos Humanos dicta una resolución²²⁵ de promoción y protección de todos los derechos humanos, civiles, políticos, económicos, sociales y culturales, incluido el derecho al desarrollo, en la cual se recuerda la universalidad, indivisibilidad, interdependencia e interrelación de todos los derechos humanos y libertades fundamentales, y afirmando que todos los derechos que se aplican fuera del entorno digital, también se aplican dentro de este entorno.

Esta Resolución también reconoce que las tecnologías digitales nuevas y emergentes tienen el potencial para facilitar los esfuerzos por acelerar el progreso humano, promover y proteger los derechos humanos y las libertades fundamentales, eliminar todas las brechas digitales, apoyar,

²²⁵ Resolución aprobada por el Consejo de Derechos Humanos el 14 de julio de 2023, A/HRC/RES/53/29.

entre otros, el disfrute de los derechos de las personas que se encuentran en situaciones de vulnerabilidad o marginación.

Además, sigue reconociendo los riesgos que las tecnologías digitales nuevas y emergentes entrañan para la protección, promoción y disfrute de los derechos humanos, incluidos entre otros, el derecho a la no discriminación y el derecho a la igualdad, de conformidad con las obligaciones de los Estados en virtud del derecho internacional de los derechos humanos.

Estas nuevas tecnologías pueden potenciar el desarrollo, garantizando siempre que todas las personas tengan acceso a ellas sin ningún tipo de discriminación.

Debemos destacar, lo rápido que cambio el mundo tecnológico, y no se pueden llegar a comprender todos los efectos que generan las nuevas tecnologías en los derechos humanos, pero es necesario seguir analizándolos de forma exhaustiva, a fin de aprovechar al máximo todo el potencial que tienen las nuevas tecnologías para apoyar el progreso y desarrollo de las personas, y más, teniendo en cuenta, si una tecnología determinada puede afectar a la situación jurídico administrativa de una persona extranjera, con los derechos humanos afectados en la toma de tal decisión. Además, para autores como MARTÍN DELGADO, la automatización de las actuaciones administrativas conlleva el respeto al derecho a la igualdad por no existir subjetividad en la toma de decisiones y evitar así la vulneración a la arbitrariedad de la Administración, garantizando la imparcialidad y la seguridad jurídica (Martín 2009).

Discriminación algorítmica en los procedimientos administrativos de extranjería

Uno de los desafíos que está suscitando el uso de aplicaciones basadas en la IA es el de la discriminación algorítmica o los resultados discriminatorios que pueden producir el uso de sistemas de decisión automatizada o semiautomatizada.

Primero, es esencial identificar en que consiste la discriminación algorítmica y segundo, cómo podemos mitigarla.

Se entiende la discriminación algorítmica como “el fenómeno que se produce cuando los sistemas pueden utilizar diferentes fuentes de datos para llevar a cabo sus decisiones, como datos históricos, patrones de comportamiento o datos de encuestas, entre otros” (Iturmendi 2023). La discriminación deviene de que estos datos pueden contener sesgos. Datos que pueden ser personales, de sexo o género, o que contienen por ejemplo la pertenencia étnica o racial de las personas, provocando la exclusión de ciertos grupos de personas como pueden ser las personas extranjeras.

Otra definición de discriminación algorítmica, sesgo de la IA o sesgo algorítmico, puede ser la existencia de errores sistemáticos y repetibles en un sistema informático que conllevan o

generan resultados discriminatorios o injustos desde un punto legal, como favorecer a un grupo de usuarios en detrimento de otros.

En base a tales definiciones es importante también destacar para una mejor comprensión de fenómeno qué es un algoritmo, teniendo en cuenta que no existe una definición universal y única, por lo que se puede definir como “un conjunto de instrucciones que un sistema informático sigue para aprender y extraer patrones a partir de los datos” (Bellio 2023).

Se ha de favorecer el acceso a la toma de decisiones mediante algoritmos para evitar en la gran medida posible que se toman decisiones basadas en discriminaciones algorítmicas violando así el derecho a la dignidad de las personas extranjeras o a su derecho a la intimidad o a la interdicción de la discriminación.

Las normativas de protección de datos personales no abarcan todos los criterios ni los mecanismos de discriminación, y tampoco establecen medidas que puedan solucionar estos casos. Pero se debe poner el foco en atribuir responsabilidad a aquellas personas que programan los algoritmos, para que en su labor no se apliquen en la toma de decisiones algoritmos sesgados.

En el desarrollo de las normas que regulen y controlen el uso de la IA, la discriminación necesita un tratamiento específico que se conecte con los mecanismos de garantía antidiscriminatorios ya existentes y previstos por la Constitución española y el Derecho. Ello evitaría áreas de exención respecto al principio de igualdad y la proliferación de sistemas basados en IA a los que se permite ignorar o violar derechos fundamentales y el principio de igualdad.

La discriminación algorítmica resulta de la interacción compleja entre el uso de las tecnologías de la información y la comunicación (TIC) y la IA y las dinámicas sociales ya existentes de desigualdad, los estereotipos y las jerarquías sociales. Abordar la discriminación algorítmica significa abordar esta complejidad (Digital Future Society 2022).

Un número creciente de estudios muestran que los sistemas de decisión automatizada o semiautomatizada basados en algoritmos, en el ámbito público, aunque también en el privado, pueden afectar a los derechos fundamentales.

El principio de eficiencia administrativo toma asiento en la toma de decisiones a través de medios automatizados, sin la intervención directa de una persona, gracias a la capacidad de almacenar una gran cantidad de datos y de procesarlos para emitir una decisión final.

Validez del acto administrativo dictado a través de una decisión automatizada

La definición clásica de acto administrativo la encontramos en Zanobini, posteriormente completada por García de Enterría, para quien el acto administrativo es “cualquier declaración

de voluntad, de juicio, de conocimiento o de deseo, emanada de un sujeto de la Administración Pública en el ejercicio de una potestad administrativa distinta de la potestad reglamentaria”. Es importante expresar el concepto de acto administrativo en vista a determinar si esta declaración de voluntad se produce cuando se dicta un acto administrativo, tal como es la resolución expresa por la que se conceden o deniegan autorizaciones de extranjería al amparo de una decisión automatizada. O lo que es lo mismo, si todo acto administrativo puede dictarse al amparo de una actuación automatizada o si por el contrario existen excepciones. Se trata de un debate mantenido por la doctrina administrativa que encuentra su fundamento en la teoría del órgano administrativo y del acto administrativo.

Si bien, autores como PARADA, considera que una decisión tomada de forma automatizada no puede considerarse resolución administrativa por cuanto “las máquinas carecen de voluntad” (Parada 1999) y uno de los contenidos del acto administrativos es expresar la voluntad de la Administración, como es conceder o denegar las autorizaciones de extranjería.

En contra, autores como MARTÍN DELGADO, consideran que detrás de la decisión automatizada están las personas físicas integrantes de la Administración que realizan materialmente la actuación. Por lo que la aplicación informática ejecuta aquello que la persona a programado y por ende la actuación se imputa al órgano que tiene encomendada la competencia (Martín 2009).

Este autor entiende por actuación administrativa automatizada “toda declaración de voluntad, juicio, conocimiento o deseo realizada por una Administración pública en el ejercicio de una potestad administrativa mediante el empleo de un sistema de información adecuadamente programado, y adaptada en el caso concreto sin necesidad de intervención de una persona física”.

Las actuaciones automatizadas se aplican tanto en actos de trámite, de comunicación o de requerimiento, y para dictar el acto por el que se resuelve el procedimiento.

La validez del acto administrativo generado de forma automatizada podría encontrar su nulidad si vulnerara el principio de motivación consagrado en la LPAC, que también tiene su acopio en el derecho a la transparencia. Toda decisión administrativa ha de ser motivada, en especial, cuando se están limitando los derechos de las personas como es el residir o residir y trabajar en España de forma legal²²⁶. Es por ello, que, para dotar las actuaciones de seguridad jurídica y

²²⁶ Así lo dispone el artículo 35.1 de la Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas.

transparencia en el caso de estar ante actuaciones automatizadas, es necesario que se determine cuándo se está ante ella, y en particular, en qué casos la terminación del procedimiento deriva de una respuesta automática, debiéndose identificar tal extremo en el propio acto administrativo.

En lo referente a los derechos de las personas cuando la Administración Pública hace uso de la IA en los procedimientos administrativos, como el procedimiento administrativo en materia de extranjería, CAPDEFERRO determina que “La Administración pública, con el objetivo último de contribuir a la adecuada consolidación de una administración inteligente, que sea eficiente, personalizada, electrónica y garante de los derechos de la ciudadanía, está incorporando, progresivamente, esta tecnología en el servicio público. Esta nueva administración va acompañada al mismo tiempo de una serie de riesgos con una clara dimensión jurídica y ética. El necesario avance tecnológico para la mejora del servicio público puede entrar, pues, en tensión con la seguridad jurídica, derechos fundamentales como la igualdad, la intimidad o la protección de datos personales; o con principios u obligaciones del derecho a defensa o recurso, pues no siempre será posible conocer el porqué de los resultados proporcionados por algunas inteligencias artificiales” (Capdeferro 2020).

Bibliografía

AÑÓN, MARÍA JOSE, “Igualdad y Procedimiento Administrativo para inmigrantes”. [en línea], [consulta:23/09/2025]. Disponible en: <https://www.uv.es/cefd/14/anyon.pdf>

BELLIO, NAIARA. *Algoritmos y desigualdad: sistemas policiales predictivos y de perfilamiento en el ámbito policial y judicial en España* [en línea]. 2023. [consulta: 09 de octubre de 2025]. Disponible en: <https://www.algorace.org/wp-content/uploads/2025/06/Informe-Algoritmos-y-desigualdad-JusticiayPolicia-ES-.pdf>.

CERRILLO MARTÍNEZ, Agustí. *Actuación automatizada, robotizada e inteligente*. Madrid: Marcial Pons, 2024. Págs. 531-547. ISBN 9788413817422.

España Digital 2026 (2021). Carta de los Derechos Digitales. ISBN 9788411472043.

España. Ley Orgánica 4/2000, de 11 de enero, sobre derechos y libertades de los extranjeros en España y su integración social. Boletín Oficial del Estado, 12 de enero de 2000, núm. 10, p. 54.

España. Real Decreto 1155/2024, de 19 de noviembre, por el que se aprueba el Reglamento de la Ley Orgánica 4/2000, de 11 de enero, sobre derechos y libertades de los extranjeros en España y su integración social. Boletín Oficial del Estado, 20 de noviembre de 2024, núm. 280, p. 178.

España. Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas. Boletín Oficial del Estado, 2 de octubre de 2015, núm. 236, p. 73.

España. Ley 40/2015, de 1 de octubre, de Régimen Jurídico del Sector Público. Boletín Oficial del Estado, 2 de octubre de 2015, núm. 236, p. 123.

GOBIERNO DE ESPAÑA. *Flujo de autorizaciones de residencia concedidas a extranjeros año 2024* [en línea].2024. [consulta: 23 de septiembre de 2025]. Disponible en: https://www.inclusion.gob.es/web/opi/estadisticas/catalogo/flujo_autorizaciones.

GOBIERNO DE ESPAÑA. *La discriminación algorítmica en España: límites y potencial del marco legal*” [en línea]. 2022: [consulta: 23 de septiembre de 2025]. Disponible en: https://digitalfuturesociety.com/app/uploads/2022/09/Discriminacion_algoritmica_Espana_marco_legal.pdf

Ministerio de Asunto Económicos y Transformación Digital. Declaración de los Derechos y Principios en el Entorno Digital. España, 2021. Disponible en: https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf

Inteligencia artificial, Derecho administrativo y reserva de humanidad algoritmos y procedimiento administrativo debido tecnológico: revista General de Derecho Administrativo. Madrid: lustel. 2002-. ISSN 1696-9650.

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, 2023. Vol. n.º 11/2023, pág. 161. ISSN 2530-4275.

La inteligencia artificial en el sector público: desarrollo y regulación de la actuación administrativa inteligente en la cuarta revolución industrial: revista de internet, derecho y política, 2020. N.º 30, pág. 2. ISSN 1699-8154.

La regulación europea de la IA ante los sesgos y riesgos de discriminación algorítmica en contextos migratorios: revista CIDOB d'Afers Internacionals. 2024. Vol. n.º 38, págs. 171-194. ISSN 11336595.

PARADA, Ramón. *Régimen jurídico de las Administraciones Públicas y Procedimiento Administrativo Común.* Madrid: Marcial Pons, 1999. ISBN 9788472487161.

Los riesgos a los derechos humanos por la inteligencia artificial: su intento de mitigación en la EU AI ACT: revista Mexicana de Derecho Constitucional. Ciudad de México: Instituto de Investigaciones Jurídicas de la Universidad Nacional Autónoma de México. 2025-. ISSN 2448-4881.

Naturaleza, concepto y régimen jurídico de la actuación administrativa automatizada: revista de Administración Pública. 2009. Vol. n.º 180, págs. 354-36. ISSN 0034-7639.

Resolución 53/29, de 14 de julio de 2023, sobre las tecnologías digitales nuevas y emergentes y los derechos humanos, p. 2-6.

Unión Europea. Directiva 2003/109/CE, del Consejo, de 25 de noviembre de 2003, relativa al estatuto de los nacionales de terceros países residentes de larga duración. Diario Oficial de la Unión Europea, 25 de noviembre de 2003, núm. L00044-00053, p. 10.

Unión Europea. Carta de los Derechos Fundamentales de la Unión Europea (2010/C 83/02). Diario Oficial de la Unión Europea, 7 de junio de 2016, núm. Z00389.00403, p. 15.

María Lourdes Soto Rodríguez. La Inteligencia Artificial y su impacto en los Derechos Humanos. Ámbito normativo y desafíos que se deben afrontar

María Lourdes Soto Rodríguez

Jueza Sustituta Adscrita al Tribunal Superior de Justicia de Galicia

Criminóloga e Investigadora.

Universidad de Santiago de Compostela.

Universidad Stanford. Course International Women's Health and Human Rights

[ACCEDE A LA GRABACIÓN](#)

Resumen

Es la era de la inteligencia artificial y es incuestionable que ha llegado para quedarse y debemos preguntarnos hasta qué punto esto juega con nuestros derechos fundamentales. En este trabajo he llevado a cabo un exhaustivo análisis normativo, jurisprudencial y doctrinal en relación con el impacto de los derechos fundamentales frente a la utilización de la inteligencia artificial. He escogido la línea Temática A, sobre “Investigaciones académicas sobre derechos humanos e inteligencia artificial: orígenes del tema, evolución hasta la actualidad, desarrollos normativos previos y propuestas para el futuro” y a través de la modalidad On line. En el presente estudio realizo un exhaustivo análisis jurídico en el que profundizo en la manera en que los gobiernos utilizan la tecnología en su propio interés y como esto repercute en los ciudadanos y en sus derechos. Dicha regulación debe ser abordada con una legislación al respecto que delimite la invasión de los derechos fundamentales y en el marco de ese proceso, se pondrán de manifiesto las consecuencias tanto legales como éticas que se derivan del uso de la inteligencia artificial.

En dicho estudio vengo también a realizar un trabajo de acercamiento normativo a nivel nacional y comunitario, enfocado en la protección de datos y los derechos fundamentales. Realizo un estudio de la jurisprudencia relevante en esta misma materia. Además procedo a identificar de manera clara los desafíos que debemos afrontar para proteger nuestros derechos en una sociedad cada vez más digitalizada y por ello concreto los múltiples desafíos que pueden aparecer en relación con los sistemas de identificación biométrica. No hay duda de que veremos avances en la tecnología de Inteligencia Artificial en los próximos años pues las tecnologías existentes ya podrían automatizar actividades laborales que ocupan entre el 60 y el 70% del tiempo de los empleados y ante ello se hace un análisis de la necesidad de explorar posibles escenarios futuros.

Capítulo I. Introducción

En un Estado de Derecho, el uso de la inteligencia artificial debe estar sujeto al principio de legalidad, lo que implica cumplir con las normas jurídicas y las regulaciones establecidas. Nos enfrentamos a varios desafíos, siendo el principal la necesidad de regular su desarrollo, mediante disposiciones legales y no simplemente basándose en consideraciones éticas. Pero evidentemente no puede dejarse de lado la perspectiva ética, pues la inteligencia artificial constituye un nuevo fenómeno que viene a revolucionar nuestras vidas.

Constituye dicho uso de la inteligencia artificial un problema para el Estado de Derecho sino para nosotros mismos, al tener que convivir con otro tipo de entes autónomos e inteligentes.

Es un hecho real la preocupación por las implicaciones que la IA puede tener para los derechos humanos y la necesidad de que la IA se desarrolle y se utilice de una manera que respete y promueva los derechos humanos está muy presente en la actualidad ²²⁷.

Sin embargo, se trata de un fenómeno relativamente reciente, teniendo en cuenta los orígenes de esta tecnología. Hasta hace pocos años, más allá de reflexión sobre la privacidad y la protección de datos, aspectos que, sin lugar a dudas, pueden abarcar espacios amplios en el campo del discurso de los derechos ²²⁸, la reflexión parecía ir más bien por el lado de la ética y no de los derechos.

En 2014 terminaba *Una mirada a la robótica desde los derechos humanos*²²⁹ 4 defendiendo la necesidad de tomar como referencia los derechos humanos a la hora de regular la robótica y, en general, lo que denominaba como tecnologías convergentes. Y es que, la presencia de los derechos humanos en la reflexión sobre el impacto en los derechos de estas tecnologías sólo estaba presente en alguna de estas tecnologías (especialmente en el ámbito de la biomedicina de la mano de la bioética) ²³⁰.

Capítulo II. Fundamentos teóricos

²²⁷ J.I. SOLAR CAYÓN y O. SÁNCHEZ MARTÍNEZ, *El impacto de la inteligencia artificial en la teoría y en la práctica jurídica*, La Ley, Madrid 2022, pp. 215 y ss; M.A. PRESNO LINERA, *Derechos fundamentales e Inteligencia Artificial*, Marcial Pons, Madrid 2023; F. J. ANSUÁTEGUI ROIG, “Nuove Tecnologie e Spazio Pubblico”, trad. A. di Rosa, en S. SALARDI y M. SAPORITTI (eds.), *Le tecnologie ‘moral’ emergente e le sfide etico-giuridiche delle nuove soggettività*, Giappichelli, Torino, 2020, pp. 22-41.

²²⁸ A. GARRIGA, *Nuevos retos para la protección de datos personales en la era del big data y de la computación ubicua*, Dykinson, Madrid 2015.

²²⁹ Dykinson, Madrid 2015.

²³⁰ Vid. V. MORENTE y C. LEMA, “Genoma humano y derechos humanos”, en G. PECESBARBA y otros, *Historia de los derechos fundamentales*, Siglo XX, vol. 4, t. 5, 2013, pp. 345 y ss.

1. Concepto de Inteligencia Artificial

El término Inteligencia Artificial (IA) hace referencia a un tipo de máquinas y sistemas con ciertas capacidades, pero existe controversia sobre cuáles son sus rasgos definitorios. Algunos autores focalizan la esencia de la IA en su capacidad de aprendizaje continuo, otros en su manera de imitar los procesos cerebrales humanos.

La inteligencia artificial puede definirse como el conjunto de prácticas, principios y herramientas científicas que buscan emular, a través de una máquina, las capacidades cognitivas de los individuos, partiendo de datos aportados por el diseñador, el usuario o recolectados por el propio dispositivo desde fuentes digitales o sensoriales del entorno físico²³¹.

Es importante destacar las características del sistema de la IA²³²:

-Adaptabilidad y aprendizaje: las IA aprenden a partir de los datos que se les aportan, adaptando su funcionamiento y volviéndose más sofisticadas y complejas con el paso del tiempo. Dentro de este denominado aprendizaje mecánico, existe el aprendizaje profundo, que describe la forma mediante la cual los sistemas de IA utilizan una red de capas superpuestas de neuronas artificiales que permiten el procesamiento simultáneo de grandes cantidades de datos, imitando la estructura y funcionamiento del cerebro humano.

- Autonomía: Una de las características más intrínsecas al propio concepto de la IA es la autonomía de sus sistemas, pues estos están diseñados para funcionar sin requerir una constante intervención humana.
- Racionalidad: El funcionamiento de la IA se basa en su capacidad de aplicar procesos lógicos a la información que se le proporciona, y formular conclusiones razonables a partir de estos.
- Percepción e Identificación: Las herramientas de IA son capaces de detectar los estímulos en su entorno e identificar patrones en conjuntos de datos. La proliferación de sistemas de IA de reconocimiento biométrico es una de las principales preocupaciones en el debate sobre el uso de la IA.

La inteligencia artificial es un proyecto de la ciencia iniciado hace 70 años pero que aloja en su seno una pulsión utópica que conecta con los orígenes utópicos y deterministas de la ciencia, concretamente con el *Nuovum Organum* de Francis Bacon y con el desarrollo que alrededor de esta idea plantea Thomas Hobbes en *De Homine*, donde se introduce la idea de que el

²³¹ Elisa Gutiérrez García, Inteligencia artificial y derechos fundamentales, 2024, p. 23.

²³² Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25, 1097-1105

conocimiento es poder y que, al ser un poder, tiene una capacidad de acción sobre el mundo que puede transformar este sobre la propia capacidad de poder que genera, y que en sí mismo tiene una energía transformadora.

En 2023, representantes de distintos estados y organizaciones internacionales como la Unión Europea se reunieron en el Reino Unido para debatir sobre las implicaciones éticas del desarrollo de la IA, produciendo la conocida como Declaración de Bletchley. Dicho documento recoge una serie de principios y postulados encaminados a propiciar una gestión responsable y colaborativa del fenómeno de la IA.

La colaboración internacional iría encaminada a asentar unas bases éticas de la IA basadas en principios como la no discriminación y la igualdad de acceso, pero, fundamentalmente, dicha base ética debería estar orientada desde el antropocentrismo, es decir, transcurra como transcurra el desarrollo de la IA, esta debe permanecer siempre como una herramienta en beneficio de la sociedad y el ser humano, no como algo que amenace su propia existencia.

Para ello es necesario encontrar el punto de equilibrio entre los beneficios aportados por la IA y la protección de los derechos de los ciudadanos, para lo que es absolutamente necesario cierto grado de supervisión e involucración humana en los procesos de IA. Es imprescindible crear un sistema de responsabilidad respecto a los actos de la IA que permita dirigirse a los desarrolladores e implementadores de esta.

la Declaración de Bletchley pone en relevancia que el desarrollo de la IA es un asunto de tal importancia que trasciende las fronteras, ofreciendo a la humanidad en su totalidad oportunidades al igual que amenazadas, y debiendo por tanto ser tratado como un asunto de vital importancia, en el que la colaboración internacional para preservar el beneficio humano es imprescindible.

José María Lassalle, en su libro *Civilización artificial*²³³, distingue dos modelos de IA:

a) Por un lado, la Nihilista, surgida en Estados Unidos y China, y diseñada para el poder sin propósito, ya sea hacia el control estatal o corporativo. En China, está vinculada al paradigma de Estado-plataforma, mientras que, en Estados Unidos, se basa en el mercado digital y el egoísmo.

²³³ Lassalle, J. M. (2018). *Civilización Artificial*. Editorial Taurus.

b) el modelo Humanista que encuentra su base en Europa, y está centrado desde una perspectiva teórica en lo humano y ético, pero que según Lassalle aún se enfrenta a desafíos para incorporar dimensiones éticas en la regulación.

2. Estado de derecho. Definición y principios fundamentales

El Estado de Derecho es una de las creaciones intelectuales más asombrosas de la humanidad. En primer lugar, supone que el mismo poder que emite las normas consiente en ser autolimitado por estas normas de manera que éstas se convierten en leyes abstractas que regulan tanto el comportamiento del propio estado como el de los ciudadanos, es decir, el Estado se autolimita. El Estado de Derecho es por tanto un principio de gobernanza que establece que tanto la sociedad civil como el propio Estado están limitados por una ley abstracta que establece los límites de su actuación. En segundo lugar, el Estado de Derecho establece la interdicción de la arbitrariedad de los poderes públicos, los cuales deben estar sujetos en su actuación a esa legislación abstracta. Dicho corpus legislativo habrá de ser público y conocido, de forma que todos tengan una previsibilidad de las consecuencias de sus actos, y deberá aplicarse en base a los principios de igualdad y no discriminación, así como de proporcionalidad.

El Estado de Derecho se configura en base a la separación de poderes, de forma que cada una de sus tres ramas podrá actuar de forma independiente dentro de sus limitadas capacidades, evitando de esta forma una acumulación excesiva de poder en ninguna institución estatal. Los poderes públicos, en sus respectivas áreas de actuación, deberán velar por el mantenimiento y aplicación de los derechos fundamentales, asegurando la dignidad y la libertad de los ciudadanos.

La calificación del Estado español como democrático de Derecho aparece consagrada en la Constitución Española de 1978, cuyo art. 1.1 dispone que *“España se constituye en un Estado social y democrático de Derecho, que propugna como valores superiores de su ordenamiento jurídico la libertad, la justicia, la igualdad y el pluralismo político”*²³⁴. Esta primera referencia actúa como piedra angular en la arquitectura constitucional, plasmando la preocupación del legislador constituyente por construir unas sólidas y estables instituciones representativas²³⁵. A su vez, en el art. 9.1 se establece que, *“Los ciudadanos y los poderes públicos están sujetos a la Constitución y al resto del ordenamiento jurídico”*²³⁶ y se complementa con la salvaguarda

²³⁴ Art. 1 CE

²³⁵ Contreras M. (1986). Iniciativa legislativa popular y estado democrático de derecho (una aproximación a su regulación jurídica en España). *Revista De Las Cortes Generales*, (8), 67-94. <https://doi.org/10.33426/rcg/1986/8/445>.

²³⁶ Art. 9 CE

del art. 103 que expresa cómo la estructura del Estado y sus entidades públicas, respaldadas por la Constitución, garantizan su protección²³⁷. Por último, en el art. 117 se subraya que la administración de justicia emana del pueblo, lo que implica la existencia de tribunales independientes sujetos exclusivamente a la ley²³⁸.

3. Intersección entre Inteligencia Artificial y estado de derecho

La democracia digital, definida como el fortalecimiento de la democracia tradicional mediante tecnologías de la información y comunicación, puede mejorar la participación ciudadana y la calidad de los procesos democráticos. Sin embargo, es importante reconocer que no garantiza la participación política por sí sola y debe complementarse con medidas fuera del ámbito digital.

Puede aplicarse en procesos donde la ciudadanía participa en la elaboración de normativas, tanto para recopilar datos e información como para evaluar la calidad de los procesos participativos. Sin embargo, es crucial abordar la desigual capacidad de intervención ciudadana que puede surgir del uso de la IA.

Por ello, es fundamental que los sistemas de IA utilizados por las Administraciones Públicas estén sujetos a control jurisdiccional para garantizar su legalidad y evitar posibles abusos, y que las decisiones administrativas basadas en la IA sean susceptibles de ser revisadas por los tribunales.

Es esencial que en el uso de la IA se respeten los derechos civiles, democráticos y sociales de los ciudadanos, lo que requiere un enfoque equilibrado que proteja tanto la eficiencia como los derechos fundamentales.

Hay que ser conscientes del peligro que puede suponer la tecnología que contiene elementos de inteligencia artificial, ya que conforme esta disciplina avanza y se consiguen herramientas más sofisticadas, estas herramientas se infiltran en nuestros sistemas de toma de decisiones, de forma que los sistemas de toma de decisiones digitales amenazan con desplazar a aquellos en los que intervienen los humanos, debido al afán de estos últimos por conseguir la mayor eficacia posible.

²³⁷ Art. 103 CE

²³⁸ Art. 117 CE

La tecnología se considera frecuentemente como una “espada de doble filo” porque sus impactos en la sociedad pueden ser tanto positivos como negativos. Por ejemplo, aunque puede limitar la libertad de expresión, también puede promoverla.

El caso Wisconsin- Loomis, en Estados Unidos ilustra algunos de los peligros que presenta la aplicación de la IA. La Administración de Justicia adoptó el algoritmo COMPASS, diseñado para, en base a datos tanto generales como específicos del caso, determinar el riesgo de reincidencia del individuo en diferentes etapas del proceso penal. Esto daba siempre como resultado un porcentaje de reincidencia más elevado en las personas de raza negra. Se rechazó en apelación, señalando cuatro motivos a tener en consideración; primero, que se trata de un software privado y por tanto carece de transparencia, al no conocerse todas las exactitudes de su funcionamiento; segundo, el sistema COMPASS evalúa el riesgo de reincidencia en grupos sociales, no en individuos concretos; tercero, utiliza datos nacionales en lugar de datos específicos de Wisconsin; y cuarto, diversos estudios demuestran que los algoritmos de sentencia tienden a clasificar desproporcionadamente a los delincuentes minoritarios como de mayor riesgo.

El artículo 63 del Reglamento General de Protección de Datos (RGDP) que amplía el artículo 22 sobre IA, otorga a los interesados el derecho a conocer y recibir información sobre la lógica detrás del procesamiento de datos en decisiones automatizadas, proporcionando una posible explicación de la tecnología. Sin embargo, este derecho se ve limitado cuando los secretos comerciales y los derechos de propiedad intelectual prevalecen sobre la transparencia. Para conseguir corregir el desequilibrio creado por la IA, se pueden emplear diversos mecanismos, como puede ser recurrir a terceros de confianza que pueden ayudar a asegurar que los algoritmos empleados se desarrollen y utilicen conforme a los valores del Estado de Derecho. En este sentido, en el Reino Unido, la Comisión de Derecho sobre el Uso de Algoritmos en el Sistema de Justicia recomendó la creación de un Registro Nacional de Sistemas Algorítmicos para verificar los algoritmos utilizados en el sistema de justicia penal.

4. Riesgos de la Inteligencia Artificial

El lanzamiento en el año 2023 de Chat GPT-4, una herramienta dotada de la capacidad de sintetizar y producir texto, ha suscitado un considerable interés acerca de cómo la inteligencia artificial generativa remodelará el ámbito laboral, el acceso a los servicios gubernamentales y la interacción en las plataformas de Internet en general (*Global*, 2023).

La inteligencia artificial generativa puede generar oportunidades, pero la ausencia de una regulación adecuada y eficaz podría agravar los riesgos para los derechos humanos en diversos campos, como por ejemplo en el acceso a la protección social, a la educación y el empleo, respecto a los derechos laborales, la privacidad y la seguridad digital, entre otros. Entre los

peligros potenciales, se encuentra el riesgo de acentuar el racismo y a aumentar el discurso de odio en Internet. Un claro ejemplo es el empleo de la tecnología de reconocimiento facial por parte de Israel en los Territorios Palestinos Ocupados, que ha aumentado las restricciones a la libertad de circulación y ha contribuido a mantener el sistema de apartheid.

En el continente americano en 2023, concretamente en Nueva York, el Departamento de Policía admitió haber utilizado esta tecnología para vigilar las protestas del movimiento Black Lives Matter en la ciudad, lo que incrementó las presiones sobre el Ayuntamiento de Nueva York para que las prohibiera.

El gobierno chino utiliza los sistemas de IA para reforzar su autoritario control sobre la población de diversas formas. Aunque el gobierno chino lleva décadas censurando a sus detractores y realizando campañas propagandísticas, es indudable que las herramientas de IA facilitan en sumo la labor de manipulación de la información y supresión de la libertad de expresión. La combinación de la extensa red de videovigilancia existente en China con las herramientas de IA de reconocimiento facial y análisis es una realidad.

Es importante tener en cuenta en el debate sobre el desarrollo de la IA el concepto de la dignidad. Ciertamente, la mayoría de las consideraciones y reglas éticas que rigen nuestra sociedad hoy en día derivan del concepto de la dignidad humana. En base a la dignidad humana, defendemos los derechos y obligaciones que el hombre debe ostentar, establecemos el límite entre lo permitido y lo prohibido, lo justo y lo injusto, lo decente y lo inhumano. En base a la dignidad humana, Kant estableció que el ser humano es un fin en sí mismo, nunca un medio, derivando de esta conceptualización gran parte de nuestro sistema de derechos fundamentales y universales. El desarrollo de la IA no solo afecta al Estado de Derecho en cuanto a leyes de protección de consumidores o de responsabilidad de las IA se refiere, sino que puede transformar por completo todas las consideraciones éticas y jurídicas en torno a las que hemos basado no solo nuestra forma de vida, sino nuestra concepción propia como especie.

Capítulo III. Desafíos y preocupaciones éticas

1. Responsabilidad legal de los sistemas de Inteligencia Artificial

La introducción de la inteligencia artificial en el mundo del Derecho conlleva aspectos positivos, pero también presenta importantes desafíos. La imposibilidad de conocer el comportamiento de la máquina ha generado la necesidad de establecer una normativa específica, ya que las reglas tradicionales no se adaptan adecuadamente a esta nueva realidad.

Surge la propuesta de un Reglamento del Parlamento Europeo y del Consejo, de 21 de abril de 2021, que establece normas armonizadas sobre inteligencia artificial (Ley de Inteligencia Artificial), aprobada preliminarmente en diciembre de 2023. Aunque no aborda el tema de la responsabilidad civil, a través de requisitos, condiciones, demandas, prohibiciones y, en

particular, mediante la clasificación de los sistemas, incorpora un marco para garantizar un uso responsable de esta tecnología.

El artículo 3.1 de la propuesta en cuestión (enmienda 165) subraya que se trata de *“un sistema basado en máquinas diseñado para funcionar con diversos niveles de autonomía y capaz, para objetivos explícitos o implícitos, de generar información de salida –como predicciones, recomendaciones o decisiones– que influya en entornos reales o virtuales”*. Algunos autores (Fernández, 2024) incluyen dentro de los subcampos de la inteligencia artificial el aprendizaje automático, el aprendizaje profundo, el procesamiento del lenguaje natural y la visión por computador.

Se establecen ciertas prohibiciones según las distintas aplicaciones de la inteligencia artificial, así como una serie de requisitos para los sistemas de inteligencia artificial considerados de alto riesgo, pudiendo resumir los aspectos más relevantes de dichas prohibiciones y requisitos en los siguientes puntos:

- Prohibiciones sobre sistemas de IA que interactúan con personas físicas: Opera la regla fundamental de que dichos sistemas deberán informar a los usuarios de que están tratando con herramientas de IA y no con personas físicas.
- Prohibiciones sobre sistemas de reconocimiento de emociones: En aras de proteger la intimidad y privacidad de los usuarios, estos sistemas se someterán a fuertes restricciones, como la prohibición de su uso en lugar de trabajo.
- Normas de seguimiento y supervisión del mercado: Deberán implantarse mecanismos de supervisión y seguimiento del uso de las IA, de forma que se asegure el cumplimiento del resto de condiciones y requisitos.

Capítulo IV. Marco regulatorio de la IA

1. Legislación sobre Inteligencia Artificial y derechos fundamentales

El artículo 9.2 de la Constitución Española establece que en un Estado democrático, es fundamental reconocer la importancia del pluralismo, el respeto a la minoría y la participación ciudadana en las distintas funciones del Estado,

La Estrategia Nacional de Inteligencia Artificial (ENIA) propone establecer un marco ético y normativo para fortalecer la protección de los derechos individuales y colectivos, promoviendo la inclusión y el bienestar social. El documento conocido como la Declaración de Toronto, suscrita el 16 de mayo de 2018, emitido por organizaciones como *Human Rights Watch*, *Access Now* y Amnistía Internacional, marca un hito en la historia al buscar aplicar los estándares internacionales de derechos humanos al desarrollo y uso de sistemas de aprendizaje automático, también llamado inteligencia artificial instando a los gobiernos y empresas a

asegurar que las aplicaciones de aprendizaje automático e inteligencia artificial se rijan por los principios de igualdad y no discriminación.

La Declaración de Toronto establece directrices de derechos humanos que tanto el sector público como el privado deben seguir para garantizar una aplicación justa y equitativa de los algoritmos, así como proporcionar vías de reparación significativas para aquellos cuyos derechos hayan sido violados. Hace un llamamiento a los gobiernos para asegurar el debido proceso en el uso de aprendizaje automático en el ámbito público, ejercer cautela en la implementación de sistemas de IA en el sistema penal, establecer mecanismos claros de rendición de cuentas en el desarrollo e implementación de aplicaciones de IA, y definir claramente quienes son los responsables legales de las decisiones tomadas mediante el uso de tales sistemas. La Declaración de Toronto, reconoce que la inteligencia artificial es una herramienta estratégica con beneficios significativos para ciudadanos, empresas y la sociedad en general, siempre y cuando se oriente hacia el ser humano, sea ética y sostenible, y respete los derechos y valores fundamentales²³⁹.

Directrices de la UE para una IA confiable

El informe de la Agencia Europea de los Derechos Fundamentales de la Unión Europea (en adelante, FRA), *Getting the future right - Artificial intelligence and fundamental rights*, afirma que el uso de los sistemas de IA siempre pondrá en peligro a una gran diversidad de derechos fundamentales, indistintamente del campo de aplicación, así, los más afectados son la privacidad, la protección de datos, la no discriminación y el acceso a la justicia (FRA, 2021, p.5). La Organización de las Naciones Unidas (en adelante, ONU) ha podido identificar la rigurosidad de la práctica, creando cuerpos exclusivos para su estudio, como el Grupo de Expertos *ad hoc* de la ONU para la Educación, la Ciencia y la Cultura (en adelante, UNESCO), el que es responsable de adoptar el primer acuerdo mundial sobre la ética de la IA¹⁹, y el Panel de alto nivel sobre cooperación digital de la Secretaría General de la Organización.

La Unión Europea es consciente de que se está abocado a un futuro sostenible y digital, donde los sistemas de IA participaran, activamente, en un sin número de ámbitos de la vida cotidiana, más de lo que ya intervienen. Esto fuerza, a los juristas a establecer un derecho capaz de comprender las necesidades nuevas y particulares que varían a máxima velocidad. Por lo tanto, desde el 2018 comenzaron los actos preparatorios para una regulación efectiva, con el Comunicado de prensa *Artificial intelligence: Commission kicks off work on marrying cutting-*

²³⁹ Santiago Camilo Carretero Sánchez, *La responsabilidad del Estado en sus nuevos frentes: sanitario, alimentario, energético y de inteligencia artificial*, 2023.

edge technology and ethical standards, el lanzamiento de la alianza europea de IA y el Plan coordinado sobre IA. Con el paso del tiempo, en abril de 2021, la Comisión Europea presentó un paquete de IA centrado en la excelencia y la confianza, conformado por un marco jurídico innovador basado en los derechos fundamentales y cuatro niveles de riesgo (Comisión Europea, 2023). En 2019, la Comisión Europea emitió una serie de directrices y recomendaciones para orientar el uso de la IA. Estas disposiciones no vinculantes adoptan un enfoque antropocentrista para describir una serie de principios éticos que aseguren un uso confiable y adecuado de la IA. Puede hablarse de siete principios clave: Intervención y supervisión humanas: Es necesaria la supervisión humana para asegurar que la IA no socaba nuestra propia autonomía ni genera efectos indeseados en nuestra persona.

En el uso de la IA deben garantizarse mecanismos de control en la aplicabilidad de los sistemas de IA, manteniendo una constante participación humana basada en el control y supervisión. Dichos mecanismos son:

a) PRIVACIDAD Y GESTIÓN DE DATOS: Para que los ciudadanos se sientan suficientemente confiados como para proporcionar a la IA sus datos personales, es fundamental garantizar una adecuada protección y gestión de estos. Las personas deben mantener en todo momento el control sobre sus datos, y ser aseguradas de que no se usarán para fines desfavorables o ilícitos. No solo deberá garantizarse que los datos están seguros ante ataques externos o manipulaciones, sino que deberán eliminarse los sesgos e inexactitudes de los conjuntos de datos empleados para entrenar a la IA, de forma que no ofrezca posteriormente resultados erróneos.

b) TRANSPARENCIA: El funcionamiento interno de los sistemas de IA deberá ser comunicado a sus usuarios, tratando en la medida de lo posible de dar luz sobre los procesos internos que conducen a los resultados aportados por la IA. Es fundamental que la actuación de la IA sea guiada por la trazabilidad, de forma que queden registradas no solo las decisiones tomadas por estas herramientas, sino los procesos que han seguido para llegar a ellas. Diversidad, no discriminación y equidad: La IA debe evitar sesgos, tanto en los datos en base a los que se educa y entrena, como en los resultados que aporta. Los sistemas de IA deberán tener en cuenta todo el espectro de capacidades y aptitudes humanas, de forma que su diseño sea lo más accesible e igualitario posible.

c) SEGURIDAD TÉCNICAS: Los sistemas de IA deben ser lo suficientemente sofisticados y desarrollados como para asegurar la integración de mecanismos de control y seguridad que aseguren tanto la veracidad de sus resultados como la protección ante ataques externos. Deberá asegurarse la mayor fiabilidad y reproducibilidad posible de sus datos, así como la reversibilidad de las consecuencias no deseadas derivadas de su funcionamiento.

1.2. La Carta de Derechos Digitales

La Carta de Derechos Digitales , presentada en 2021, y resulta ser una directriz no vinculante encaminada a describir las situaciones de conflicto dadas entre el actual desarrollo tecnológico y la protección de los derechos individuales, para ofrecer así un marco ético y legislativo que permita consensuar la implantación de nuevas tecnologías como la IA en nuestra sociedad con la preservación de aquellos derechos, valores y principios derivados de la dignidad humana que sustentan el Estado de Derecho.

La carta recoge una serie de derechos proyectables en diversas áreas como los derechos laborales o la investigación científica, y en su sección vigesimoquinta, establece una serie de derechos ante la IA orientados desde la protección de la persona y su dignidad y la búsqueda del bien común. La Carta de Derechos Digitales constituye un recordatorio directo de los principios y valores éticos que orientan la legislación española, el cual insta a los poderes públicos a dirigir la actividad legislativa en lo concerniente al desarrollo tecnológico desde el puro antropocentrismo, recordando que las nuevas herramientas tecnológicas deben servir y proteger al hombre y su dignidad.

1.3. El Real Decreto 729/2023, publicado en el BOE el 2 de septiembre de 2023, establece el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial (AESIA), que entra en vigor el 3 de septiembre de 2023.

Este estatuto aprueba la creación de la AESIA, el primer organismo de su tipo en Europa, en anticipación a la implementación del Reglamento Europeo de Inteligencia Artificial (Real Decreto 729/2023, de 22 de agosto, por el que se aprueba el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial, 2023). La agencia tiene la responsabilidad de supervisar y evaluar el desarrollo y uso de la IA en diversos sectores, asegurando que estos se alineen con los principios éticos y legales vigentes que conforman nuestro Estado de Derecho. La agencia evalúa el impacto de la IA en la sociedad y en los derechos fundamentales de los ciudadanos. En el ámbito regulatorio, la AESIA tiene la tarea de proponer y desarrollar normativas específicas para la implementación y uso de la IAE, garantizando que estas promuevan la seguridad, la transparencia y el respeto a los derechos humanos. La AESIA suministra asesoramiento técnico a entidades públicas y privadas, promoviendo buenas prácticas y el cumplimiento de las regulaciones establecidas.

1.4. El Real Decreto 817/2023, publicado en el BOE el 9 de noviembre de 2023, establece un entorno controlado de pruebas para evaluar el cumplimiento de la propuesta de Reglamento del Parlamento Europeo y del Consejo sobre normas armonizadas en inteligencia artificial, que entra en vigor el 10 de noviembre de 2023.

Esta normativa pone en marcha el primer entorno controlado de pruebas para verificar cómo implementar los requisitos aplicables a los sistemas de Inteligencia Artificial de alto riesgo según la propuesta de Reglamento Europeo. Esta iniciativa se integra en la estrategia española

de transformación digital, conocida como Agenda España Digital 2026, dentro del Plan de Recuperación y, específicamente, en la Estrategia Nacional de Inteligencia Artificial (Real Decreto 817/2023, de 8 de noviembre, que establece un entorno controlado de pruebas para el ensayo del cumplimiento de la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial, 2023). El objetivo de la Propuesta de Reglamento del Parlamento Europeo y del Consejo que establece normas armonizadas sobre Inteligencia Artificial (Acta de Inteligencia Artificial) y modifica ciertos actos legislativos de la Unión, es abordar las deficiencias de las reglas existentes, así como establecer un sistema de control efectivo para los sistemas de IA. Esta propuesta reglamentaria establece un enfoque basado en el riesgo y establece cuatro niveles de riesgo:

- Sistemas cuyo riesgo es tan inaceptablemente alto que están prohibidos (Título II)
- Sistemas que generan alto riesgo (Título III)
- Sistemas a los que, aunque no se consideren de alto riesgo, se aplican una serie de requisitos de transparencia (Título IV)
- Los sistemas restantes (Título IX)

1.5. El Reglamento Europeo de Inteligencia artificial (AIA, Artificial Intelligence Act)

El 21 de abril de 2024, el Consejo de la Unión Europea dio luz verde a la Ley de Inteligencia Artificial.

La Ley de la IA prohíbe:

- Técnicas subliminales, manipulativas o engañosas que alteren el comportamiento de personas o grupos, llevándolos a decisiones que normalmente no tomarían y que probablemente les causen daño significativo.
- Explorar las vulnerabilidades de personas o grupos debido a su edad, discapacidad o situación socioeconómica, para distorsionar su comportamiento de manera que probablemente les cause daño significativo.
- Sistemas de IA que predicen delitos basándose únicamente en el perfil o los rasgos de personalidad de una persona.
- Sistemas de IA que expanden bases de datos de reconocimiento facial mediante la recopilación no autorizada de imágenes fáciles de Internet o CCTV.
- Sistemas de IA que infieren emociones en lugares de trabajo y educativos, salvo por razones médicas o de seguridad.
- Sistemas de IA que categorizan a personas basándose en datos biométricos para inferir datos sensibles.

- Sistemas de identificación biométrica remota en tiempo real en espacios públicos para la aplicación de la ley, con ciertas excepciones y garantías.

2 . Legislación nacional y jurisprudencia sobre Inteligencia Artificial en España

2.1. Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.

Según la Constitución Española el tratamiento de datos personales es considerado un derecho fundamental por el Art.18.4 que dispone que “la ley limitara el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos” (art. 18.4 CE).

Por otro lado, la Sentencia 94/1998, del 4 de mayo, hace referencia a la Sentencia 254/1993 que asume la declaración de que el Art.18.4 de la Constitución España añade una protección adicional a los derechos y las libertades fundamentales, en particular, en relación con el derecho al honor y la intimidad; la interpretación jurisprudencial señala lo siguiente: la llamada libertad informática es así derecho a controlar el uso de los mismos datos insertos en un programa informático, *habeas data*⁵², y comprende, entre otros aspectos, la oposición del ciudadano a que determinados datos personales sean utilizados para fines distintos de aquel legítimo que justifico su obtención (STC 94/1998, de 4 de mayo, FJ. 4º). A continuación, la sentencia hace referencia a la normativa que desarrolla lo establecido en el Art.18.4 de la Constitución Española y destaca, como un principio cardinal de la protección de datos, la necesidad de que su uso sea congruente y racional, con una protección reforzada de los datos sensibles. Además, la sentencia prohíbe taxativamente, de manera explícita, el uso de estos datos para fines distintos a los que motivaron su obtención legítima y reconoce una tutela especial para los datos sensibles (STC 94/1998, de 4 de mayo, FJ. 4º). Estos puntos se adaptan al ordenamiento jurídico en la Ley Orgánica 3/2018, del 5 de diciembre, de Protección de datos personales y garantía de los derechos digitales⁵³ (en adelante, LOPDGDD), al igual que el RGPD. Conforme con dicha ley, el derecho fundamental recogido en el Art.18.4 se ejercerá con arreglo al RGPD y garantizara los derechos digitales de la ciudadanía que nacen de dicha protección (Art.1, LOPDGDD).

Es significativo señalar los artículos (en adelante, Arts.) que protegen los puntos de mayor relevancia, como el Art.9 que hace referencia a las categorías especiales de datos, como los biométricos. En este apartado, a efectos del Art.9.2.a) del RGPD, se refuerza la protección al objeto de evitar situaciones discriminatorias, “el solo consentimiento del afectado no bastara para levantar la prohibición del tratamiento de datos cuya finalidad principal sea identificar su ideología, afiliación sindical, religión, orientación sexual, creencias u origen racial o étnico”(Art.9.1, LOPDGDD). Por ello, el tratamiento de dichos datos quedara bajo el amparo de los restantes supuestos contemplados en el Art.9.2 del RGPD (Art.9.1, LOPDGDD) de estos; es

relevante mencionar el apartado g) por razones de un interés público esencial y el i) por razones de interés público en el ámbito de la salud pública.

2.2. Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación.

Otra disposición legal de fecha reciente y la primera regulación positiva del uso de la IA por las administraciones públicas y las empresas en España ²⁴⁰ es la Ley 15/2022 del 12 de julio, integral esta para la igualdad de trato y la no discriminación (Ley 15/2022). Según se está ante una regulación programática y voluntarista que elabora las líneas de actuación de las administraciones públicas, para poder favorecer, promover y priorizar aquellas políticas y prácticas ligadas con el uso de algoritmos involucrados en la toma de decisiones. En el marco de esta normativa, cabe resaltar el Art.23, este se refiere específicamente a la IA y los sistemas de toma de decisiones automatizadas, asimismo, se menciona que, dentro de la perspectiva del marco de la Estrategia Nacional de IA de la Carta de Derechos Digitales y las iniciativas europeas en torno a la IA, la Administración Pública española deberá fomentar la implementación de medidas que permitan que los algoritmos utilizados en la toma de decisiones en las instituciones públicas consideren criterios de minimización de sesgos, transparencia y rendición de cuentas, siempre que sea viable desde una perspectiva técnica (art.23.1, Ley 15/2022). Dichas medidas incluirán el diseño de los algoritmos y los datos utilizados para entrenarlos, así, deberán abordar su probable repercusión discriminatoria. Para alcanzar este último objetivo, se llevaran a cabo evaluaciones de impacto que identifiquen cualquier sesgo discriminatorio potencial (art.23.1, Ley 15/2022), de igual modo, se exige a las administraciones públicas y a las empresas promover el empleo de una IA “ética, confiable y respetuosa con los derechos fundamentales, siguiendo especialmente las recomendaciones de la Unión Europea en este sentido” (art.23.2, Ley 15/2022), lo que se verificara a través de un sello de calidad de los algoritmos (art.23.4, Ley 15/2022).

3. Jurisprudencia destacable

La jurisprudencia consolidada hace referencia a conflictos de esta índole en todos los aspectos, por las actuaciones de las entidades privadas y el de las autoridades públicas, por ello, se consideran relevantes aquellas en el ámbito comunitario que tratan sobre dicha cuestión, para comprender el nivel de sensibilidad de la utilidad/recopilación de datos biométricos. Algunas a destacar son aquellas del Tribunal de Justicia de la Unión Europea (en adelante, TJUE), como la Sentencia de 13 de mayo de 2014, *Google Spain, S.L. y Google Inc. contra AEPD y Mario Costeja*

²⁴⁰ María Angeles Guervos, “Fiscalidad de las smart cities” 2022, p.49

Gonzalez y la Sentencia de 6 de octubre de 2015, *Data Protection Commissioner* contra *Facebook Ireland Limited* y Maximillian Schrems.

A su vez, el TEDH ha creado un valioso apoyo jurisprudencial para la data biométrica, en especial, respecto con la obtención de muestras celulares, perfiles del ADN, huellas dactilares, huellas palmares y muestras de voz, lo que se encuentra desarrollado en la *Guide to the Case-Law of the European Court of Human Rights, Data protection, Genetic and biometric data* (TEDH, 2022, pp.12-14). Analizar el papel que cumple el Estado para la protección de estos derechos es primordial, de este modo, gracias a la jurisprudencia, se evidencia desde en la *praxis* cuales son las exigencias. Por lo tanto, estos poderes de control y disposición sobre los datos personales, que forman parte del contenido del derecho fundamental a la protección de datos, se concretan jurídicamente en la capacidad de dar consentimiento a “la recogida, la obtención y el acceso a los datos personales, su posterior almacenamiento y tratamiento, así como su uso o usos posibles, por un tercero, sea el Estado o un particular” (STC 292/2000, de 4 de enero, FJ. 7º).

En el siguiente supuesto, la jurisprudencia fija una serie de estándares sobre el tratamiento de datos biométricos, concretamente, los adquiridos de la huella fácil de los perjudicados. Indiferentemente de que este tratamiento se realice por una entidad privada, el lugar de su aplicación es considerado un espacio de acceso público conforme con la Ley de IA, por lo que requieren las mismas prohibiciones si fuera realizado por las autoridades públicas.

Mencionar el Auto de la Audiencia Provincial de Barcelona 1448/2021, Sección 9ª, Auto 72/2021 de 15 Feb. 2021, Rec. 840/2021 en la que la sociedad mercantil, Mercadona S.A., presentaba el 6 de mayo de 2019 un escrito que manifestaba la intención de implementar un sistema de vigilancia en sus supermercados, para poder realizar el reconocimiento facial de aquellas personas con sentencias firmes y órdenes de alejamiento en vigor contra Mercadona o sus empleados. Dicha solicitud fue denegada por el Juzgado de lo penal núm. 24 Barcelona en el auto en 33.fecha 27 de septiembre de 2019, a efectos de su desproporcionalidad, razón por la cual en fecha 25 de noviembre de 2020, la mercantil presenta recurso de apelación resuelta en la Sentencia de la Audiencia Provincial (en adelante, SSAP) de Barcelona 1448/202158. A causa del aprovechamiento de este sistema piloto de reconocimiento facial en múltiples sucursales, el cual inicio el 1 de julio de 2020 y se practicó durante varios meses, se presentaron dos reclamaciones, una el día 15 de julio de 2020, numero de registro 025103/2020, de la Asociación de Consumidores y Usuarios en Acción-Facua, otra el 27 de julio de 2020, numero de registro 026511/2020, procedente de Apedanica, por lo que inicio el procedimiento sancionador de AEPD PS/00120/2021. Tanto en la revisión del tribunal como en el procedimiento sancionador, se afirma que la aplicación de esta IA es un tratamiento desproporcionado. A simple vista, resulta un tratamiento excesivo, debido a que para hacer

eficaz una medida de seguridad para un número limitado de personas, en este caso solo dos personas, por un periodo limitado y establecido en sentencia, que no podrá exceder de seis meses al tratarse de un delito leve (art. 57.3 del Código Penal), se podrían monitorear a un colectivo de más de 100.000 trabajadores y 1.624 establecimientos para la fecha (AEPD, 2021, p.84).

El Mercadona argumenta que no existe un tratamiento de datos, ya que, lleva a cabo la identificación en tiempo real y elimina en 0,3 segundos toda la información, solo utilizando resultados positivos para contactar a las autoridades en caso de detección. Esto dentro del análisis es insignificante, en palabras del tribunal, “resulta, no obstante, cuanto menos sorprendente que se amparen en la "rapidez". Por muy rápido que sea, existe una violación de la privacidad. Tanto el argumento de la rapidez como el no tratamiento de datos caen por su propio peso” (SSAP 1448/2021, de 15 de febrero, FJ. 2º). A efectos de su alta complejidad, la sentencia realiza un acercamiento doctrinal a la AEPD en su informe 36/2020. Gracias a su interpretación, se reafirma que el uso de tecnologías de reconocimiento facial en sistemas de videovigilancia implica el tratamiento de datos biométricos que tienen como objetivo identificar de forma única a una persona física, este tipo de tratamiento de datos constituye una categoría especial que, en principio, está prohibido por el art. 9.1 del RGPD. Se establece de manera explícita que “para tratar categorías especiales de datos con estos fines, la normativa requiere que exista un interés público esencial recogido en una norma con rango de ley que no existe actualmente en el ordenamiento jurídico” (SSAP 1448/2021, de 15 de febrero, FJ. 2º). En otras palabras, para que el reconocimiento facial tenga las mínimas garantías, debe existir una protección legal específica, la cual no existe en nuestro ordenamiento jurídico. Se establece que el Mercadona a través de este proyecto piloto ha infringido un número significativo de normas, como el art.6 del RGPD relativo a la licitud del tratamiento de datos personales, el art.5.1.c) referente del principio de minimización de datos y el art.9 anteriormente expuesto. Por consiguiente, se desestima el recurso de apelación, se confirma íntegramente el Auto de fecha 27 de septiembre de 2019, dictado por el Juzgado de lo Penal N.º 24 de Barcelona y se le sanciona con el pago de 3,15 millones de euros, los cuales luego fueron reducidos a 2.520.00 € por pago voluntario.

Capítulo V. Estrategias de regulación y gobernanza de la Inteligencia Artificial en el ámbito legal
Estamos ante nueva generación de normativas de protección de datos, destacando el

Reglamento General de Protección de Datos (RGPD) de la Unión Europea del año 2016²⁴¹. Estas regulaciones se han adoptado progresivamente en diversos países, destacándose por su carácter proactivo y preventivo.

Principios:

- a) El principio de responsabilidad proactiva, requiere que el responsable del tratamiento no solo c cumpla con la normativa de protección de datos, sino que también sea capaz de demostrar su cumplimiento.²⁴²
- b) Gestión del riesgo, que implica la evaluación del riesgo para determinar las medidas de seguridad adecuadas según el nivel de riesgo identificado²⁴³.

Estos principios conllevan la obligación de incorporar la protección de datos desde la etapa de diseño y por defecto, lo que implica la aplicación de medidas adecuadas considerando el estado actual de la tecnología, los costes involucrados y la naturaleza del tratamiento de datos, así como los riesgos para los derechos y libertades de los individuos afectados²⁴⁴. Además de estas medidas, se incluyen otras acciones complementarias que promueven la responsabilidad activa, como la designación de un delegado de protección de datos y los mecanismos de autorregulación a través de códigos de conducta y certificaciones, lo que evidencia que la normativa de protección de datos se encuentra a la vanguardia de este nuevo enfoque de gobernanza del riesgo en el ámbito digital, estableciendo estándares que abarcan todo este dominio. No se puede afirmar que en el ámbito de la innovación, el principio de precaución haya sido desplazado por el principio de innovación, aunque este último haya emergido recientemente y carezca, por el momento, de reconocimiento formal en el Derecho europeo o nacional, limitándose su referencia a documentos informales que lo definen como un “impulso a la regulación que fomente la innovación manteniendo los estándares de protección de derechos y valores reconocidos por la Unión”²⁴⁵.

²⁴¹ 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 (DOU 4.5.2016 L 119/1)

²⁴² Art. 5 RGPD

²⁴³ Art. 35 RGPD

²⁴⁴ Art. 25 RGPD

²⁴⁵ European Political Strategy Centre (EPSC), «Towards an Innovation Principle Endorsed by Better Regulation», EPSC Strategic Notes nº 14, 30 de junio de 2014.

Capítulo VI. Conclusiones

Es obvio que la IA está transformando de manera vertiginosa diversos aspectos de la sociedad y la economía, incluido el ámbito jurídico. Debemos aspirar a un entorno digital seguro, que garantice la protección de la privacidad y los datos personales. La IA plantea desafíos a mayores en la democracia digital, especialmente en lo que respecta al control de plataformas tecnológicas dominantes y la manipulación de datos personales para influir en las decisiones de los votantes. Por lo tanto, se requieren estrictas obligaciones de transparencia y medidas efectivas de cumplimiento para mitigar estos riesgos. El desarrollo de la IA tiene el potencial para alterar de forma radical conceptos tales como la conciencia, la humanidad, la dignidad y la consciencia, por lo que su implementación tiene un impacto directo en el Estado de Derecho. Las implicaciones de la Inteligencia Artificial en el Estado de Derecho en España son de una magnitud considerable, y su comprensión y abordaje requieren una atención meticulosa y multidisciplinaria. En España, como en muchas jurisdicciones, el desarrollo y la implementación de tecnologías basadas en el uso de la IA plantean desafíos y oportunidades significativas para el Estado de Derecho. La inteligencia artificial puede optimizar la eficiencia y la accesibilidad de la justicia, agilizando los procedimientos judiciales no obstante existe la preocupación de intromisiones y alteraciones con la privacidad, la equidad y la transparencia, que deben ser abordadas para garantizar la salvaguardia de los derechos fundamentales de los ciudadanos.

El marco legal actual en España ofrece ciertas protecciones y salvaguardias en relación con el uso de la IA en el ámbito jurídico, aunque es imprescindible adaptarlo y fortalecerlo para hacer frente a los desafíos emergentes. Para ello es preciso establecer normativas claras y específicas que regulen el desarrollo, la implementación y el uso de sistemas de IA en el sistema judicial. Dichas normativas deben fijar estándares éticos y técnicos rigurosos, así como mecanismos de supervisión y control, para garantizar que la IA sea utilizada de manera justa, transparente y responsable, y que ésta constituya en todo caso una herramienta destinada a beneficiar al ser humano. Se hace necesario proporcionar educación y capacitación en materia de tecnología y derecho, promoviendo el diálogo público y la participación ciudadana en la formulación de políticas relacionadas con la IA. Se precisa una colaboración estrecha entre los diferentes actores involucrados en el desarrollo y la implementación de la IA en el ámbito jurídico, incluidos los poderes públicos, el sector privado, la academia y la sociedad civil. Solo a través de un enfoque colaborativo y multidisciplinario se pueden abordar de manera efectiva los desafíos complejos que plantea la IA en el Estado de Derecho. Esto incluye la promoción de la investigación interdisciplinaria, la cooperación internacional y el intercambio de mejores prácticas y lecciones aprendidas.

A mi juicio es preciso adoptar un enfoque crítico y reflexivo hacia la IA, evaluando cuidadosamente sus beneficios y riesgos en cada contexto específico y asegurando que se utilice de manera complementaria y subordinada a los principios y valores fundamentales del

Estado de Derecho. El diseño y la implantación de los sistemas de IA deberá ajustarse siempre a principios como el de transparencia, no discriminación, supervisión humana y rendición de cuentas, para asegurar que los ciudadanos son conscientes en todo momento del funcionamiento de las herramientas a las que se someten, así como para garantizar que no sufren un trato discriminatorio y velar por su derecho a revertir los efectos lesivos derivados del incorrecto funcionamiento o uso de la IA.

Bibliografía

A. GARRIGA, Nuevos retos para la protección de datos personales en la era del big data y de la computación ubicua, Dykinson, Madrid 2015.

Contreras M. (1986). Iniciativa legislativa popular y estado democrático de derecho (una aproximación a su regulación jurídica en España). Revista De Las Cortes Generales, (8), 67-94. <https://doi.org/10.33426/rcg/1986/8/445>.

Elisa Gutiérrez García, Inteligencia artificial y derechos fundamentales, 2024.

Guervos, M. (2022). Fiscalidad de las Smart Cities (1.a ed.). ARANZADI.

J.I. SOLAR CAYÓN y O. SÁNCHEZ MARTÍNEZ, El impacto de la inteligencia artificial en la teoría y en la práctica jurídica, La Ley, Madrid 2022, pp. 215 y ss.; M.A. PRESNO LINERA, Derechos fundamentales e Inteligencia Artificial, Marcial Pons, Madrid 2023; F. J. ANSUÁTEGUI ROIG, “Nuove Tecnologie e Spazio Pubblico”, trad. A. di Rosa, en S. SALARDI y M. SAPORITTI (eds.), Le tecnologie ‘moralì’ emergente e le sfide etico-giuridiche delle nuove soggettività, Giappichelli, Torino, 2020.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25, 1097-1105.

Lassalle, J. M. (2018). Civilización Artificial. Editorial Taurus.

Santiago Camilo Carretero Sánchez, La responsabilidad del Estado en sus nuevos frentes: sanitario, alimentario, energético y de inteligencia artificial, 2023.

V. MORENTE y C. LEMA, “Genoma humano y derechos humanos”, en G. PECESBARBA y otros, Historia de los derechos fundamentales, Siglo XX, vol. 4, t. 5, 2013.

Legislación

Constitución Española (BOE 29 de diciembre de 1978).

Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación.

La Carta de Derechos Digitales.

Real Decreto 729/2023, publicado en el BOE el 2 de septiembre de 2023, establece el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial (AESIA), que entra en vigor el 3 de septiembre de 2023.

Agencia por los Derechos Fundamentales de la Unión Europea (FRA).

Reglamento General de Protección de Datos (RGPD).

Real Decreto 817/2023, publicado en el BOE el 9 de noviembre de 2023, establece un entorno controlado de pruebas para evaluar el cumplimiento de la propuesta de Reglamento del Parlamento Europeo y del Consejo sobre normas armonizadas en inteligencia artificial, que entra en vigor el 10 de noviembre de 2023.

Reglamento Europeo de Inteligencia artificial (AIA, Artificial Intelligence Act) .

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 (DOU 4.5.2016 L 119/1).

Reglamento (UE) 2021/0106 del Parlamento Europeo y del Consejo de 21 de abril de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión.

Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.

La Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD).

Listado de abreviaturas

Agencia Española de Supervisión de Inteligencia Artificial (AESIA).

Agencia Española de Supervisión de Inteligencia Artificial (AESIA).

Artificial Intelligence Act (AIA).

Estrategia Nacional de Inteligencia Artificial (ENIA)

Inteligencia Artificial (IA).

Reglamento General de Protección de Datos (RGPD).

Unión Europea (UE).

Mercedes Ten Doménech. Automatización algorítmica y Derechos Humanos: Un llamamiento a una gobernanza ética global

[ACCEDE A LA GRABACIÓN](#)

Resumen

En el contexto de la acelerada expansión tecnológica, la relación entre inteligencia artificial (IA), derechos humanos y democracia se configura como una cuestión central en la agenda global. La IA interviene de manera creciente en ámbitos decisivos —desde la gestión pública hasta la selección laboral o la seguridad—, con la capacidad de fortalecer o debilitar los principios de inclusión, equidad y participación que sustentan toda democracia. El sesgo algorítmico constituye uno de los principales desafíos. Al reproducir prejuicios presentes en los datos históricos o en los diseños de los modelos, los sistemas de IA pueden generar decisiones discriminatorias, opacas e inapelables, que afectan de manera desproporcionada a los grupos en situación de vulnerabilidad: mujeres, migrantes, minorías étnicas, personas mayores o con discapacidad. Este fenómeno no es meramente técnico, sino político y ético, pues compromete la igualdad ante la ley, la rendición de cuentas y la legitimidad democrática. Los impactos del sesgo algorítmico se manifiestan en sectores clave como la sanidad, las finanzas, el empleo, la seguridad y la gobernanza digital, ampliando desigualdades estructurales y erosionando la confianza ciudadana. Además, la concentración de poder tecnológico en grandes corporaciones y potencias globales ha generado una nueva forma de colonialismo digital, que amenaza la soberanía de los Estados y la capacidad colectiva de decidir el propio futuro. En este escenario, la gobernanza democrática de la IA resulta indispensable. Solo un marco normativo robusto, transparente y orientado a la dignidad humana puede asegurar que la inteligencia artificial sea una aliada de la democracia y no un instrumento de exclusión o control.

Introducción

Uno de los rasgos distintivos del orden jurídico y ético contemporáneo es el reconocimiento de que toda persona, por el simple hecho de pertenecer a la especie humana, posee derechos fundamentales que no pueden ser legítimamente negados por ninguna autoridad. Estos derechos no emanan de la voluntad del Estado, ni constituyen una concesión suya; tampoco están condicionados por la nacionalidad, la etnia, la religión o cualquier otra característica cultural. Se trata de derechos de carácter universal, imprescriptibles e inalienables. Esta conquista se ve reflejada de manera paradigmática en el artículo 1º de la Declaración Universal de Derechos Humanos, el cual proclama que “todos los seres humanos nacen libres e iguales en dignidad y derechos y, dotados como están de razón y conciencia, deben comportarse fraternalmente los unos con los otros”.

En este marco, el desarrollo y uso de la inteligencia artificial (IA) plantea una serie de interrogantes cruciales en relación con la vigencia y el respeto de estos derechos humanos (DD. HH). Al tratarse de tecnologías que, cada vez con mayor autonomía, intervienen en decisiones que afectan directamente a las personas —desde la asignación de recursos públicos hasta la evaluación de perfiles laborales, pasando por mecanismos de vigilancia o control social—, es imprescindible garantizar que su implementación no menoscabe la dignidad humana ni reproduzca estructuras de exclusión, discriminación o injusticia. Así mismo lo afirmaba la Alta Comisionada de las Naciones Unidas para los Derechos Humanos, Michelle Bachelet en septiembre de 2021: “la inteligencia artificial puede ser una fuerza para el bien, que ayude a las sociedades a superar algunos de los mayores retos de nuestro tiempo. Pero las tecnologías de IA también pueden tener efectos nocivos e incluso catastróficos, cuando se emplean sin prestar la debida atención a su capacidad de vulnerar los derechos humanos”²⁴⁶. La universalidad de los DD.HH exige que los sistemas de IA sean diseñados y regulados bajo principios que aseguren la equidad, la transparencia, la rendición de cuentas y el respeto por la autonomía individual. Así, cualquier innovación tecnológica debe estar subordinada a un marco normativo que reconozca, no solo la centralidad del ser humano, sino también la inalienabilidad de los derechos que le corresponden en virtud de su condición, sin importar su origen, estatus legal o entorno sociocultural.

²⁴⁶ *Declaración de la Alta Comisionada sobre los riesgos de la inteligencia artificial para la privacidad en el Consejo de Europa*. 2021. Disponible en: <https://www.ohchr.org/es/press-releases/2021/09/artificial-intelligence-risks-privacy-demand-urgent-action-bachelet> [Consulta: 25 de junio de 2025]

La complejidad de esta relación exige repensar los marcos tradicionales de protección jurídica en función de nuevos desafíos, entre ellos: la opacidad de los algoritmos, el riesgo de discriminación automatizada, la vigilancia masiva, la manipulación de la información y la erosión de la autonomía individual. La IA, si bien posee un potencial transformador en favor del desarrollo humano, también puede consolidar o incluso agravar desigualdades preexistentes si no es diseñada, aplicada y regulada bajo los principios de justicia, equidad y transparencia. En consecuencia, resulta necesario realizar un análisis multidisciplinario que contemple un desarrollo tecnológico respetuoso con la dignidad inherente a toda persona. La preocupación acerca del impacto de las tecnologías inteligentes en los derechos humanos comenzó a finales del siglo XX, particularmente desde la ética de la tecnología y la filosofía política y jurídica. Por aquel entonces, se intuía que los avances en automatización, vigilancia digital y toma de decisiones asistida por máquinas podían representar un riesgo importante para las libertades civiles, la privacidad y la igualdad. Sin embargo, estos planteamientos iniciales —a menudo considerados como advertencias futuristas o especulativas— no adquirieron relevancia normativa y política hasta la segunda década del siglo XXI, cuando el desarrollo exponencial de la inteligencia artificial (IA) hace tangibles muchos de los peligros previamente anticipados. En este contexto, el presente trabajo se propone examinar los factores determinantes que han favorecido la consolidación de un espacio de reflexión académica en la intersección entre la inteligencia artificial (IA) y los derechos humanos, así como analizar las principales iniciativas legislativas de alcance internacional, en atención al carácter universal de los derechos que se busca salvaguardar. A través del esclarecimiento del desarrollo histórico-intelectual y de los contextos sociojurídicos que dieron lugar a este campo interdisciplinario, se pretende comprender la legitimidad, relevancia y proyección futura de dicha articulación. El objetivo último es contribuir a una comprensión crítica del modo en que la relación entre tecnología e instituciones jurídicas ha pasado a constituirse como una preocupación prioritaria tanto para el pensamiento académico como para la formulación de políticas públicas y marcos normativos a nivel global.

1. El imparcial sistema de decisión automatizada

Origen del sesgo algorítmico.

Uno de los aspectos más relevantes y urgentes dentro de este ámbito de reflexión interdisciplinaria en auge es la identificación de los riesgos que ciertas aplicaciones de la inteligencia artificial plantean para la protección efectiva de los derechos humanos. Entre ellos, el sesgo algorítmico ha adquirido una visibilidad particular, no sólo por su incidencia directa en decisiones que afectan a individuos y colectivos vulnerables, sino también por el modo en que desafía los principios de igualdad, no discriminación y el derecho a un proceso justo. Este fenómeno, que se encuentra en la intersección entre lo técnico y lo jurídico, constituye un

punto de inflexión en los debates contemporáneos sobre la legitimidad del uso de sistemas automatizados en contextos sensibles.

En términos generales, el sesgo algorítmico alude a la producción de resultados sistemáticamente desfavorables —o, en algunos casos, injustificadamente favorables— hacia determinados individuos o grupos, como resultado de deficiencias estructurales en el diseño de los modelos o en los conjuntos de datos que los alimentan. Aunque los algoritmos son concebidos, al menos en su diseño inicial, como instrumentos de objetividad técnica y neutralidad operativa, en la práctica pueden reproducir —e incluso amplificar— desigualdades sociales preexistentes, lo que conduce a decisiones discriminatorias y contrarias a los principios de equidad y justicia. Lo cierto es que el surgimiento del sesgo algorítmico responde a una variedad de causas interrelacionadas, que abarcan desde el diseño de los sistemas hasta las características de los datos empleados.

Uno de principales promotores de sesgo algorítmico son los propios datos que alimentan los sistemas y entrenan los modelos. Dado que los algoritmos de aprendizaje automático se nutren de grandes volúmenes de datos históricos, cualquier prejuicio presente en dichos datos — como sesgos raciales, de género o socioeconómicos— es susceptible de ser internalizado por el sistema y proyectado en sus decisiones futuras. Así, por ejemplo, un modelo entrenado con datos de contratación laboral que reflejan una histórica preferencia por candidatos masculinos tenderá a replicar dicho patrón en futuras selecciones, perpetuando la desigualdad de género²⁴⁷. Por otra parte, el diseño del algoritmo en sí mismo puede introducir sesgos, ya sea de forma consciente o inadvertida, como resultado de las elecciones realizadas por los desarrolladores. Estas decisiones incluyen la selección de variables, la definición de los objetivos de optimización o las formas de segmentación poblacional. Por ejemplo, un sistema que prioriza la procedencia educativa de ciertos centros de prestigio podría favorecer a personas provenientes de contextos más privilegiados, excluyendo de forma indirecta a candidatos igualmente cualificados de entornos más desfavorecidos²⁴⁸.

La insuficiente representación de determinados grupos en los datos de entrenamiento es otra fuente crítica de sesgo. Cuando el conjunto de datos no refleja adecuadamente la diversidad de la población a la que se aplicará el sistema, los algoritmos tienden a producir resultados menos precisos o erróneos para los grupos subrepresentados. Este problema ha sido

²⁴⁷ BOLUKBASI, T. et al., “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, *Proceedings of the 30th International Conference on Neural Information Processing Systems, Boston University, 2016*, pp. 4349–4357.

²⁴⁸ O’NEIL, C., *Armas de destrucción matemática. ¿Cómo el Big Data alimentará la desigualdad y la amenaza democrática?*, Capitán Swing Libros, Madrid, 2018.

ampliamente documentado en tecnologías como el reconocimiento facial, donde los sistemas entrenados mayoritariamente con imágenes de personas blancas presentan tasas de error significativamente más altas al identificar a personas con tonos de piel más oscuros²⁴⁹. La falta de diversidad en los datos no solo reduce la eficacia del sistema, sino que compromete su legitimidad social y legal. Finalmente, los circuitos de retroalimentación (*feedback loops*) también pueden contribuir a la persistencia y agravamiento del sesgo algorítmico. Estos se producen cuando los resultados de un algoritmo influyen en el entorno del que extrae sus datos, reforzando patrones preexistentes. Un ejemplo típico es el de los algoritmos de recomendación en plataformas digitales, que tienden a ofrecer a los usuarios contenido similar al que han consumido previamente. Esta lógica, si no se controla, puede dar lugar a burbujas informativas que limitan la exposición a perspectivas diversas y profundizan la polarización ideológica²⁵⁰.

En conjunto, estos mecanismos evidencian que el sesgo algorítmico no es un accidente técnico aislado, sino una expresión de dinámicas sociales, históricas y epistémicas que se trasladan al ámbito digital a través de modelos matemáticos aparentemente neutrales. Lejos de constituir una anomalía corregible exclusivamente mediante ajustes técnicos, el sesgo revela las formas en que estructuras de poder, prejuicios sociales y lógicas de exclusión se codifican en sistemas tecnológicos. En este sentido, a medida que la sociedad adquiere una mayor comprensión de cómo funcionan estos sistemas, diversas organizaciones e investigadores han identificado ejemplos destacados de sesgo algorítmico en una amplia gama de aplicaciones, lo que ha reforzado la necesidad de un enfoque ético y normativo en su desarrollo y aplicación.

Repercusión de la imparcialidad algorítmica en ámbitos especialmente sensibles.

Tras haber analizado la génesis del sesgo algorítmico, resulta pertinente examinar cómo este fenómeno se manifiesta en ámbitos especialmente sensibles, donde sus consecuencias pueden tener un impacto directo y pernicioso sobre la vida de las personas. Una de estas áreas es la sanidad, en el que un modelo que utilice el gasto sanitario como indicador de necesidad médica puede, inadvertidamente, perpetuar desigualdades históricas en el acceso a la atención, o puede llevar a tratamientos basados en criterios de eficiencia, y no en dignidad humana²⁵¹. Asimismo, los datos clínicos utilizados para entrenar algoritmos predictivos suelen estar infrarrepresentados en lo que respecta a mujeres o grupos étnicos minoritarios. Este

²⁴⁹ BUOLAMWINI, J., GEBRU, T., "Gender Shades: Interseccional Accuracy Disparities in Commercial Gender Classification", *Proceedings of Machine Learning Research*, 81;1-15, 2018, pp. 1-15.

²⁵⁰ PARISER, E., *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press, New York, 2011.

²⁵¹ CASTELLANOS, J. "Human Rights, Vulnerability and Artificial Intelligence: An Analysis in Constitutional Perspective", *Deusto Journal of Human Rights*, n.º 14, 2024, p. 43

desequilibrio ha derivado en errores de diagnóstico o tratamientos inadecuados. Por ejemplo, se ha documentado que algunos sistemas de diagnóstico asistido por ordenador arrojaron resultados de menor precisión para pacientes negros que para pacientes blancos, debido a la falta de diversidad en los datos clínicos utilizados. Un estudio reveló que, para un mismo nivel de riesgo según el algoritmo, los pacientes negros presentaban enfermedades más graves, ya que el sistema asociaba gastos médicos más bajos con menor necesidad de atención. Esta discrepancia respondió a que el modelo utilizaba “costes sanitarios en lugar de enfermedad como proxy de necesidad médica” y no tuvo en cuenta que los pacientes negros reciben menos recursos, incluso cuando tienen niveles de enfermedad equivalentes²⁵².

En el ámbito del empleo, se han observado sesgos significativos en los sistemas de seguimiento de candidatos ²⁵³, que emplean algoritmos de procesamiento del lenguaje natural. Un caso emblemático es el de Amazon, que abandonó un sistema de selección automatizada tras descubrir que favorecía a candidatos masculinos. El modelo otorgaba mayor valor a términos como “contratado” o “captado”, que aparecían con mayor frecuencia en los currículos de los hombres, perpetuando así un sesgo de género estructural²⁵⁴. En el campo de la publicidad en línea, también se han identificado sesgos algorítmicos que reproducen estereotipos sociales. Un estudio de la Universidad Carnegie Mellon reveló que el sistema de anuncios de Google mostraba con más frecuencia ofertas de empleo mejor remuneradas a los hombres que a las mujeres, pese a que ambos grupos compartían perfiles similares. Además, los resultados cuantitativos fueron impactantes: los perfiles masculinos mostraron anuncios de empleos ejecutivos con altos salarios aproximadamente 1 800 veces, frente a solo 300 incidencias en los perfiles femeninos. Esto pone en evidencia cómo los algoritmos pueden reforzar desigualdades estructurales en el acceso a oportunidades laborales²⁵⁵.

El sesgo también ha sido detectado en los sistemas generativos, como es el caso de la herramienta de IA Midjourney. Una investigación académica reveló que, al generar imágenes de personas en profesiones especializadas, la aplicación representaba tanto a personas jóvenes

²⁵² OBERMEYER, Z.; POWERS, B.; VOGELI, C.; MULLAINATHAN, S. “Dissecting racial bias in an algorithm used to manage the health of populations”. *Science*, vol. 366, n.º 6464, 25 oct. 2019, pp. 447-453. DOI: 10.1126/science.aax2342.

²⁵³ CORRALES, José M. “Derecho al trabajo y discriminación: las políticas públicas de empleo como instrumento de igualdad”, en Elena García-Antón (dir.), *Los derechos humanos en la inteligencia artificial: su integración en los ODS de la Agenda 2023*. Cizur Menor: Thomson Reuters Aranzadi, 2022, p. 418.

²⁵⁴ DASTIN, J., “Amazon scraps secret AI recruiting tool that showed bias against women”, *Reuters*, 10 oct. 2018. Disponible en: <https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/>

²⁵⁵ DATTA, A.; TSCHANTZ, M. C.; DATTA, Am., “Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination”, *Proceedings on Privacy Enhancing Technologies*, vol. 2015, n.º 1, 2015, pp. 92–112. DOI:10.1515/popets-2015-0007.

como mayores, pero asignaba sistemáticamente los rostros envejecidos al género masculino. Este patrón contribuye a reforzar estereotipos de género en torno a la autoridad, la experiencia y la visibilidad de las mujeres en el ámbito profesional. Como señala el informe, “las mujeres tienden a ser representadas con una apariencia más juvenil, incluso en profesiones que implican experiencia, mientras que los hombres aparecen con signos visibles de envejecimiento asociados a autoridad y liderazgo”²⁵⁶.

Esta problemática también se extiende al ámbito de la seguridad pública y la justicia penal, donde las herramientas de policía predictiva emplean datos históricos sobre detenciones para determinar zonas de riesgo. Estos sistemas, lejos de neutralizar patrones discriminatorios, tienden a consolidarlos, generando ciclos de vigilancia excesiva sobre comunidades minoritarias y reforzando perfiles raciales en la acción policial²⁵⁷. Estas herramientas no sólo suponen un riesgo para la privacidad individual, sino que también plantean dilemas sobre la proporcionalidad, la presunción de inocencia y el derecho a no ser objeto de discriminación arbitraria. Lo cierto es que la incorporación de la inteligencia artificial (IA) en el ámbito de la seguridad y la defensa representa uno de los desarrollos más transformadores, pero también más controvertidos, de la revolución tecnológica contemporánea. En un contexto caracterizado por amenazas cada vez más complejas y conflictos de naturaleza híbrida, la IA emerge como un componente estratégico clave para el diseño y ejecución de operaciones militares, así como para la prevención y respuesta frente a riesgos emergentes. Sin embargo, este potencial conlleva también una serie de implicaciones éticas, legales y geopolíticas que exigen una reflexión crítica y regulaciones claras para preservar el equilibrio entre innovación tecnológica y seguridad internacional²⁵⁸. Uno de los principales dilemas éticos radica en la delegación de funciones críticas a sistemas autónomos, especialmente en lo relativo a la selección y neutralización de objetivos. La ausencia de empatía y juicio moral plantea serias limitaciones para la toma de decisiones compatibles con el derecho internacional humanitario²⁵⁹. Diversas voces en la comunidad científica y organizaciones como la campaña *Stop Killer Robots* han solicitado la prohibición preventiva de armas letales autónomas²⁶⁰. Se destaca la necesidad de

²⁵⁶ ZHOU, M.; ABHISHEK, V.; KANNAN, S.; et al. *Bias in Generative AI*, University of British Columbia, Universidad de California, Irvine y Carnegie Mellon University, mar. 2024. Disponible en: https://www.andrew.cmu.edu/user/ales/cib/bias_in_gen_ai.pdf

²⁵⁷ LUM, K.; ISAAC, W., “To predict and serve? Predictive policing and the ethics of preemption”, *Significance*, vol. 13, n.º 3, 2016, pp. 14-19.

²⁵⁸ UNIDIR (United Nations Institute for Disarmament Research), *The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence*. Geneva, 2021. Disponible en: <https://www.unidir.org>

²⁵⁹ SHARKEY, N., *The Ethics of Autonomous Weapons Systems*, ne KATINA, P. et al. (eds.), *The Ethics of Artificial Intelligence*. Springer, 2019. p. 153–169. ISBN 978-3-030-21834-6.

²⁶⁰ ICRAC (International Committee for Robot Arms Control), *Stop Killer Robots Campaign Briefing Paper*. 2022. Disponible en: <https://www.stopkillerrobots.org>

mantener un control humano significativo sobre los sistemas de IA, especialmente en escenarios en los que las decisiones tengan consecuencias letales.

De hecho, en términos geopolíticos, la IA se ha convertido en un vector clave del poder estatal en el siglo XXI. Tal como afirmó el presidente ruso Vladimir Putin en 2017, “quien domine la inteligencia artificial dominará el mundo”. Estados Unidos y China lideran esta competencia estratégica, consolidando su influencia global a través del desarrollo masivo de IA en sectores militares, industriales y civiles²⁶¹. En cambio, Europa enfrenta el riesgo de convertirse en un espacio dependiente tecnológicamente, situación que algunos analistas describen como “cibervasallaje” o “colonialismo digital”²⁶². Este nuevo orden digital está moldeado no sólo por Estados, sino también por corporaciones tecnológicas transnacionales como Google, Amazon, Meta, Microsoft, Alibaba o Tencent. Estas empresas a parte de controlar infraestructuras digitales críticas, también dominan el procesamiento de datos e implementación de algoritmos a escala global. Su influencia no se limita al mercado: afecta decisiones políticas, inversiones públicas y el rumbo del desarrollo científico y tecnológico en múltiples regiones del mundo.

Frente a este panorama, la comunidad académica y profesional ha subrayado la necesidad de establecer **mecanismos robustos de gobernanza de la IA**. Este concepto alude a la capacidad institucional de una organización para dirigir, supervisar y controlar las actividades relacionadas con la IA, mediante un conjunto estructurado de políticas, principios éticos y marcos técnicos. Una gobernanza efectiva no solo mitiga riesgos, sino que también permite equilibrar los beneficios de la IA para empresas, usuarios y la sociedad en su conjunto. Como señalan Floridi y colaboradores, “transparency and accountability are crucial for building public trust, as they enable individuals to understand and influence the decisions that affect their lives”²⁶³.

2. Respuestas normativas: Iniciativas regulatorias e instrumentos éticos

El desarrollo de marcos éticos y normativos en torno a la inteligencia artificial refleja una creciente toma de conciencia global sobre los profundos impactos que estas tecnologías pueden tener sobre los derechos humanos, la justicia social y el tejido democrático. Documentos como el informe del ACNUDH, la Recomendación de la UNESCO, la Recomendación del Consejo de la OCDE y la Declaración de Toronto comparten una

²⁶¹ ALLISON, G., *Destined for War: Can America and China Escape Thucydides's Trap?*, Houghton Mifflin Harcourt, Boston, 2017. ISBN 978-1328915382.

²⁶² MOROZOV, E., *Digital Democracy, Analogue Politics: How the Internet Era is Transforming Kenya*. Zed Books, Londo, 2019. ISBN 978-1786992101.

²⁶³ FLORIDI, L. et al. “AI People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations”. *Minds and Machines*, vol. 28, 2018, pp. 689–707. doi:10.1007/s11023-018-9482-5.

preocupación común: la necesidad de situar la dignidad humana, la equidad y la rendición de cuentas en el centro de la innovación tecnológica. Estos instrumentos, aunque distintos en su alcance y naturaleza jurídica, coinciden en la urgencia de promover sistemas de IA transparentes, explicables, inclusivos y sometidos a controles efectivos, tanto por parte de los Estados como del sector privado. Asimismo, evidencian la insuficiencia de los marcos regulatorios tradicionales para afrontar los desafíos emergentes de la automatización y el aprendizaje automático, abogando por la creación de nuevas herramientas normativas, institucionales y metodológicas que garanticen una gobernanza democrática de la IA. En conjunto, estos textos configuran una arquitectura incipiente de principios que, si bien aún carecen en muchos casos de fuerza vinculante, están contribuyendo a moldear el debate internacional y las políticas públicas sobre inteligencia artificial. Su convergencia en torno a valores comunes constituye un paso decisivo hacia una inteligencia artificial que no solo sea técnicamente eficaz, sino legítima, socialmente responsable y compatible con los estándares universales de derechos humanos.

2.1. Ámbito internacional

2.1.1. La Declaración de Toronto

Las preocupaciones éticas en torno al desarrollo y uso de la inteligencia artificial (IA) constituyen un eje central del debate contemporáneo en torno a las tecnologías emergentes, de modo que organizaciones internacionales han desarrollado **directrices y marcos normativos orientados a mitigar los riesgos asociados al desarrollo y uso de sistemas de inteligencia artificial (IA)** en un intento de compatibilizar la innovación tecnológica con la protección de los derechos fundamentales, la equidad social y la sostenibilidad democrática. Estas inquietudes se han intensificado particularmente en relación con el empleo de herramientas de *Big Data* y los riesgos políticos asociados a su utilización maliciosa, especialmente cuando dichas tecnologías se emplean para manipular la opinión pública o distorsionar procesos sociales y democráticos. En respuesta a ello, organizaciones de derechos humanos y tecnología —entre ellas *Human Rights Watch*, *Access Now*, Amnistía Internacional, *Canadian Civil Liberties Association*, *Open Society Foundations* o *International Human Rights Program* (University of Toronto Faculty of Law)— promueven en 2018 la Declaración de Toronto sobre la protección del derecho a la igualdad y la no discriminación en los sistemas de aprendizaje automático²⁶⁴. Este documento es el resultado de un proceso participativo que incluyó extensas consultas con decenas de expertos en inteligencia

²⁶⁴ *The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems*. 16 de mayo de 2018. Disponible en: <https://www.accessnow.org/the-toronto-declaration/> [Consulta: 25 de junio de 2025].

artificial, representantes de empresas tecnológicas, académicos y organizaciones de la sociedad civil comprometidas con los derechos humanos. La Declaración fue adoptada oficialmente el 16 de mayo de 2018 durante la conferencia *RightsCon*, y desde entonces se ha convertido en una referencia fundamental para el desarrollo ético de tecnologías de aprendizaje automático (*machine learning*).

El texto se estructura en **tres secciones principales**, orientadas a delimitar responsabilidades estatales, obligaciones empresariales y derechos de las personas afectadas por sistemas automatizados. En primer lugar, el documento establece el deber de los Estados de prevenir cualquier forma de discriminación en el diseño, desarrollo e implementación de sistemas de aprendizaje automático, tanto en el ámbito de las instituciones públicas como a través de asociaciones público-privadas. Esta sección incluye principios fundamentales como la identificación de riesgos, la obligación de garantizar **transparencia y rendición de cuentas**, la **divulgación pública del uso de sistemas automatizados**, la **supervisión independiente**, y la **promoción activa de la igualdad de trato**. En el plano internacional, la Declaración invita a incorporar estos principios en las agendas diplomáticas y propone avanzar hacia una visión compartida sobre los límites y las posibilidades del desarrollo tecnológico. De este modo, se plantea la necesidad de actualizar el corpus del derecho internacional de los derechos humanos a los desafíos emergentes de la IA. Los Estados, en tanto actores clave de la gobernanza global de estas tecnologías, están llamados a desarrollar mecanismos eficaces para asegurar la transparencia, la responsabilidad institucional y la reparación de posibles violaciones. La segunda sección de la Declaración se refiere a las **responsabilidades del sector privado**, especialmente de las empresas que desarrollan o implementan sistemas de aprendizaje automático. Estas responsabilidades se articulan a partir del marco de **diligencia debida en derechos humanos**, consagrado en los *Principios Rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos* (2011). Se exige a las empresas que identifiquen y evalúen los riesgos de discriminación, adopten medidas efectivas para prevenir y mitigar dichos riesgos —incluyendo auditorías externas e independientes en contextos de alto riesgo—, y actúen con la máxima **transparencia**. Esta última implica la publicación de especificaciones técnicas, muestras representativas de los datos de entrenamiento, y el señalamiento claro de las fuentes de información utilizadas. Finalmente, la tercera sección reafirma el **derecho a un recurso efectivo y a la responsabilización jurídica de quienes violen derechos fundamentales mediante el uso de IA**. Se insta a los Estados a garantizar un proceso justo en el uso del aprendizaje automático en el sector público, a actuar con extrema cautela en la aplicación de estos sistemas en el ámbito penal, a definir claramente las líneas de rendición de cuentas y a establecer quiénes —organismos o individuos— deben responder legalmente por las decisiones automatizadas. Esta dimensión refuerza la idea de que la IA no puede operar al margen del principio de responsabilidad jurídica, ni sustituir la estructura de garantías que rige el Estado de derecho. Como enfatiza la Declaración en su cierre, los avances tecnológicos no deben

socavar nuestros derechos humanos [...] quienes tienen poder deben actuar ahora para protegerlos”²⁶⁵.

Aunque la Declaración se centra en los riesgos que plantea el aprendizaje automático para el derecho a la igualdad, muchos de sus principios poseen un alcance más amplio y pueden aplicarse a otros ámbitos de la inteligencia artificial, así como a un conjunto diverso de derechos fundamentales. Entre ellos se incluyen el derecho a la privacidad, la libertad de expresión, la participación en la vida cultural, el acceso a mecanismos efectivos de reparación e, incluso, el derecho a la vida, todos ellos susceptibles de verse comprometidos por decisiones automatizadas. Los promotores del texto subrayan que su publicación representa tan solo un punto de partida. Para que sus principios se traduzcan en prácticas concretas, es necesario desarrollar guías de implementación operativa que orienten el trabajo cotidiano de ingenieros, diseñadores, responsables institucionales y otros actores involucrados. La integración sustantiva del enfoque de derechos humanos en la ética de la IA y en la gestión de datos exige no solo compromiso normativo, sino también la definición de metodologías, estándares técnicos y estructuras de gobernanza sólidas.

2.1.2. Recomendación del Consejo sobre la Inteligencia Artificial

Otro de los documentos clave en la configuración de una gobernanza ética de la inteligencia artificial es la Recomendación del Consejo sobre la Inteligencia Artificial²⁶⁶, adoptada por la Organización para la Cooperación y el Desarrollo Económicos (OCDE) en mayo de 2019. Considerado el primer instrumento intergubernamental sobre IA de carácter global, este marco normativo fue adoptado por todos los Estados miembros de la organización, así como por diversas economías no miembros, consolidándose como un referente normativo en la materia. Su enfoque parte del reconocimiento de que la inteligencia artificial tiene el potencial de generar beneficios sustanciales para el crecimiento económico, el bienestar social y la sostenibilidad, pero también conlleva riesgos considerables que requieren atención específica desde el plano ético, legal y técnico.

La Recomendación promueve el desarrollo y uso de sistemas de IA que sean robustos, seguros, responsables y alineados con los valores democráticos y los derechos fundamentales. Se articula alrededor de una visión de la inteligencia artificial como herramienta al servicio del interés general, con capacidad para mejorar la vida de las personas y fomentar la prosperidad global, siempre que se garantice el respeto al Estado de derecho, la equidad, la inclusión y la

²⁶⁵ *Ibid*, p. 16.

²⁶⁶ OCDE, *Recomendación del Consejo sobre la inteligencia artificial, Traducción no oficial, París, 2019. Disponible en: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>*

diversidad. “Los sistemas de IA deben beneficiar a las personas y al planeta fomentando el crecimiento inclusivo, el desarrollo sostenible y el bienestar”²⁶⁷, se afirma en el documento. Asimismo, la Recomendación hace especial énfasis en la importancia de la transparencia, la explicabilidad y la trazabilidad de los sistemas automatizados. Se reconoce que, para que exista confianza pública en estas tecnologías, es necesario que los procesos de toma de decisiones algorítmicas puedan ser comprendidos, evaluados y supervisados por terceros, especialmente en contextos de alto impacto. Tal como establece el texto, los actores implicados “deberían proporcionar información significativa [...] para permitir a los afectados por un sistema de IA comprender los resultados”²⁶⁸. Esta transparencia debe ser acompañada de mecanismos sólidos de rendición de cuentas, de modo que las responsabilidades por los efectos de los sistemas se asignen claramente a lo largo de todo su ciclo de vida. En su dimensión operativa, el documento incluye recomendaciones prácticas dirigidas a los gobiernos, tales como la promoción de entornos de experimentación controlada (*sandboxes*), la inversión en capacidades humanas e institucionales, y el desarrollo de métricas para evaluar el impacto social y económico de la IA. Se fomenta también la cooperación internacional para evitar enfoques fragmentados y para garantizar que las ventajas de la inteligencia artificial sean compartidas globalmente, en particular con países con menores capacidades tecnológicas.

Aunque su naturaleza no es jurídicamente vinculante, la influencia de esta Recomendación ha sido notable. Ha orientado el diseño de estrategias nacionales de inteligencia artificial, inspirado normativas sectoriales y contribuido a la consolidación de un lenguaje común sobre la necesidad de una IA ética, confiable y centrada en el ser humano. Su adopción por parte del G20, y su eco en iniciativas de la Unión Europea y de la UNESCO, dan cuenta del peso normativo y político que ha alcanzado este instrumento, que posiciona a la OCDE como un actor central en el debate global sobre la gobernanza de la inteligencia artificial.

2.1.3. Recomendación sobre la Ética de la Inteligencia Artificial

Dos años más tarde, en noviembre de 2021, la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) adoptó por consenso su Recomendación sobre la Ética de la Inteligencia Artificial²⁶⁹, convirtiéndose en el primer instrumento normativo global de alcance universal en esta materia. Este documento no solo articula valores y principios rectores, sino que también formula directrices concretas de política pública, con énfasis en la

²⁶⁷ *Ibid*, secc. 1.1.

²⁶⁸ *Ibid*, secc.1.3.

²⁶⁹ UNESCO. *Recomendación sobre la Ética de la Inteligencia Artificial*. París: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, 2021. Disponible en línea: <https://unesdoc.unesco.org/ark:/48223/pf0000381133> [Consulta: 20 de junio de 2025].

inclusión, la sostenibilidad y el respeto a los derechos humanos. A diferencia de otros marcos regulatorios centrados en la eficiencia técnica o la mitigación de riesgos, la Recomendación de la UNESCO incorpora una dimensión ética y epistemológica explícita, en la que subraya la necesidad de concebir la inteligencia artificial (IA) desde una perspectiva pluralista, justa y orientada al bien común.

La Recomendación reconoce que “las tecnologías de la IA pueden ser de gran utilidad para la humanidad [...] pero también suscitan preocupaciones éticas fundamentales”²⁷⁰. Entre ellas, la presencia de los sesgos, la amenaza a la diversidad y la concentración del poder informativo. En ese marco, la UNESCO propone una visión de la ética como “una reflexión normativa sistemática [...] que puede guiar a las sociedades a la hora de afrontar de manera responsable los efectos conocidos y desconocidos de las tecnologías de la IA”²⁷¹. Esta aproximación integral abarca valores como la dignidad humana, la justicia social, la diversidad cultural, el respeto a los derechos humanos, y la protección del medio ambiente y los ecosistemas. De igual modo, la Recomendación pone especial énfasis en la necesidad de garantizar la supervisión humana, la transparencia, la explicabilidad y la rendición de cuentas a lo largo de todo el ciclo de vida de los sistemas de IA. “Los sistemas de IA deberían estar sujetos a un seguimiento exhaustivo por las partes interesadas pertinentes [...] y contar con mecanismos de evaluación del impacto ético, transparencia y trazabilidad”²⁷². Además, exhorta a los Estados Miembros a implementar estrategias nacionales en ética de la IA y a establecer mecanismos inclusivos y multidisciplinarios de gobernanza, que aseguren la participación de todos los sectores sociales: “los mecanismos de gobernanza de la IA deben ser inclusivos, transparentes, multidisciplinarios y multilaterales [...] para reducir la posibilidad de usos indebidos y garantizar beneficios compartidos”²⁷³.

2.1.4. El derecho a la privacidad en la era digital

En septiembre de 2022, la Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos (ACNUDH) publicó el informe *El derecho a la privacidad en la era digital*²⁷⁴ presentado ante el Consejo de Derechos Humanos de las Naciones Unidas. Este documento

²⁷⁰ *Ibid*, párr. 62.

²⁷¹ *Ibid*, párr. 65.

²⁷² *Ibid*, párr. 12 y 53.

²⁷³ *Ibid*, párr. 54.

²⁷⁴ OFICINA DEL ALTO COMISIONADO DE LAS NACIONES UNIDAS PARA LOS DERECHOS HUMANOS. *El derecho a la privacidad en la era digital: Informe de la Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos. Consejo de Derechos Humanos, 51.º período de sesiones, temas 2 y 3. Documento de la ONU A/HRC/51/17 [en línea]. Ginebra: Naciones Unidas, 4 ago 2022. Disponible en: <https://undocs.org/es/A/HRC/51/17>*

examina las amenazas contemporáneas al derecho a la privacidad, identificando como ejes principales de preocupación el uso de herramientas de intrusión informática, las restricciones al cifrado y la expansión de tecnologías de vigilancia en espacios públicos. Asimismo, ofrece un análisis detallado sobre el impacto de las tecnologías emergentes —en particular, la inteligencia artificial (IA)— en el ejercicio y la garantía de diversos derechos humanos. El informe pone especial énfasis en la elaboración automatizada de perfiles, la toma de decisiones algorítmica y otros sistemas de aprendizaje automático, los cuales inciden de manera directa en derechos como la salud, la educación, la libertad de circulación, reunión, asociación y expresión.

Se advierte que la creciente adopción de estos mecanismos plantea riesgos significativos para la autonomía individual y la dignidad humana, especialmente cuando los individuos no cuentan con mecanismos efectivos de conocimiento, control o impugnación frente a decisiones automatizadas. A este respecto, el informe afirma que “la digitalización de las decisiones estatales y del sector privado, si no se regula cuidadosamente, puede conducir a la arbitrariedad y la discriminación”²⁷⁵. La problemática se intensifica cuando estos sistemas se alimentan de datos incompletos, sesgados o deficientemente supervisados, lo que puede “consolidar estructuras de exclusión y marginación ya existentes”²⁷⁶. En este sentido, el informe sostiene que la opacidad que caracteriza a muchas de estas tecnologías compromete no solo la rendición de cuentas, sino también la posibilidad de obtener reparación efectiva: “la falta de transparencia y explicabilidad en los procesos algorítmicos obstaculiza la capacidad de las personas afectadas para comprender, impugnar o corregir decisiones que les afectan profundamente”²⁷⁷. A este respecto, se destaca que “las decisiones adoptadas por sistemas automatizados pueden tener consecuencias profundas en los derechos humanos. La ausencia de transparencia, la dificultad para impugnar decisiones y la falta de rendición de cuentas socavan la dignidad humana y la autonomía individual”²⁷⁸. Frente a este panorama, el Alto Comisionado concluye que los marcos normativos vigentes son “insuficientes para mitigar eficazmente los riesgos de los sistemas digitales complejos y garantizar la protección de los derechos humanos”²⁷⁹. Por ello, aboga por el fortalecimiento de principios rectores como la explicabilidad, la supervisión humana, el acceso significativo a la información y la responsabilidad institucional. En definitiva, se subraya que “los avances tecnológicos no deben

²⁷⁵ *Ibid.*, párr.51.

²⁷⁶ *Ibid.*, párr.52.

²⁷⁷ *Ibid.*, párr.53.

²⁷⁸ *Ibid.*, párr.51.

²⁷⁹ *Ibid.*, párr.55.

desarrollarse al margen del derecho internacional de los derechos humanos, sino que este debe servir como guía normativa central para su diseño, aplicación y supervisión”²⁸⁰.

2.2. Ámbito europeo

2.2.1. AI ACT.

Uno de los documentos de referencia más influyentes en el debate actual sobre la regulación de la inteligencia artificial en el contexto europeo es el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial)²⁸¹, aprobado el 13 de junio de 2024. Este instrumento normativo establece un conjunto de reglas armonizadas para garantizar que el desarrollo y uso de los sistemas de inteligencia artificial (IA) dentro del mercado único europeo se realice de forma segura, ética y conforme a los derechos fundamentales. Representa la primera legislación integral sobre inteligencia artificial a nivel mundial, situando a la Unión Europea a la vanguardia de la gobernanza digital y tecnológica.

El enfoque adoptado por la AI Act está eminentemente basado en el riesgo, lo que permite modular el grado de intervención normativa según el potencial impacto de cada sistema de IA. El reglamento clasifica los sistemas en cuatro niveles de riesgo: **inaceptable, alto, limitado y mínimo**. Los sistemas considerados de riesgo inaceptable están prohibidos por su carácter intrínsecamente lesivo para los derechos fundamentales o la dignidad humana. Este grupo incluye, por ejemplo, herramientas de vigilancia masiva sin autorización judicial, técnicas de manipulación subliminal que causen perjuicios sustanciales, o sistemas de puntuación social gestionados por autoridades públicas (art.5). Los sistemas de alto riesgo, por su parte, están sujetos a un régimen de obligaciones reforzadas. Este grupo abarca tecnologías aplicadas en ámbitos sensibles como la sanidad, la educación, el empleo, la justicia o la gestión de infraestructuras críticas. Entre los requisitos establecidos para estos sistemas se incluyen: evaluaciones previas de conformidad, controles de calidad, documentación técnica detallada, trazabilidad de datos, intervención humana en los procesos automatizados y supervisión postcomercialización (arts. 9 y ss). Estas disposiciones buscan asegurar que la innovación

²⁸⁰ *Ibid.*, párr.6.

²⁸¹ 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 se modifican determinados actos legislativos de la Unión (Ley de Inteligencia Artificial). Diario Oficial de la Unión Europea, L, 12.07.2024. Disponible en: <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32024R1689>

tecnológica no suponga un retroceso en materia de derechos fundamentales ni un riesgo para la seguridad de los ciudadanos. En el caso de los sistemas clasificados como de riesgo limitado, se exige un nivel de transparencia mínimo, como el deber de informar al usuario de que está interactuando con una inteligencia artificial. Finalmente, los sistemas de **riesgo mínimo**, que constituyen la mayoría de los desarrollos actuales, no están sujetos a obligaciones regulatorias específicas, aunque se fomenta la autorregulación mediante códigos de conducta voluntarios.

Uno de los pilares estructurales del AI Act es su fuerte compromiso con la **transparencia, la trazabilidad y la rendición de cuentas**. El reglamento establece la necesidad de que los sistemas de IA sean comprensibles y auditables, en especial aquellos que inciden en decisiones que afectan los derechos de las personas. Asimismo, contempla la obligación de notificar incidentes graves, llevar registros accesibles y cooperar con las autoridades competentes. Este enfoque busca no solo reducir los riesgos asociados a la opacidad algorítmica, sino también **fortalecer la confianza pública en las tecnologías emergentes**. El marco de gobernanza diseñado por el AI Act también merece especial atención. Se prevé la creación de un **Comité Europeo de Inteligencia Artificial**, encargado de coordinar la aplicación uniforme del reglamento, así como la designación de **autoridades nacionales competentes** responsables de la supervisión y el cumplimiento normativo en cada Estado miembro. Este entramado institucional refuerza la coherencia en la implementación y permite una vigilancia eficaz sobre los operadores tecnológicos. Además, el reglamento introduce **medidas de apoyo a la innovación**, como los llamados *regulatory sandboxes* o entornos de prueba regulados, que permitirán a startups y pymes experimentar con tecnologías de IA bajo condiciones controladas y con cierto grado de flexibilidad jurídica (art. 53). Este equilibrio entre control y estímulo a la creatividad técnica refleja la ambición de Europa de combinar liderazgo ético con competitividad digital.

En definitiva, la AI Act representa un **hito en la regulación internacional de la inteligencia artificial**, no sólo por su contenido técnico y jurídico, sino por su capacidad para proyectar un modelo normativo replicable en otras jurisdicciones. La Unión Europea reafirma así su papel como potencia normativa (*normative power*), capaz de establecer estándares globales en un ámbito estratégico para el futuro de las sociedades democráticas. Lejos de limitar la innovación, el reglamento busca canalizarla hacia un desarrollo responsable, justo y centrado en el bienestar humano.

Conclusiones

La relación entre inteligencia artificial y derechos humanos constituye una de las problemáticas más urgentes de la agenda tecnológica contemporánea. Si bien el campo ha avanzado en profundidad y diversidad, la brecha entre el desarrollo técnico y el resguardo ético y legal persiste. La consolidación de una gobernanza centrada en la dignidad humana requiere

enfoques interdisciplinarios, normativas coherentes y un compromiso sostenido con la justicia social en la era digital.

La revelación de sesgos algorítmicos en sistemas de decisión automatizada, especialmente en áreas como la predicción del comportamiento delictivo, la selección de personal, la concesión de créditos o la moderación de contenidos, ha puesto en evidencia que estos sesgos, lejos de ser excepciones, han demostrado ser sistemáticos y estructurales. Los algoritmos no solo replican, sino que muchas veces amplifican prejuicios sociales preexistentes incorporados en los datos de entrenamiento. En este sentido, el sesgo algorítmico no constituye un mero fallo técnico puntual, sino que refleja la proyección de dinámicas sociales, históricas y epistémicas en el entorno digital, a través de modelos computacionales que se presentan como neutros. Su comprensión y corrección exigen, por tanto, un enfoque interdisciplinario que articule saberes técnicos con una perspectiva crítica basada en los derechos humanos y los principios de justicia social.

Las consecuencias derivadas de la imparcialidad algorítmica adquieren una gravedad aún mayor cuando se delegan funciones decisorias a la IA en ámbitos especialmente sensibles, como la administración de justicia, la seguridad pública, la sanidad, la gestión migratoria, el empleo y las políticas de bienestar social. En estos contextos, surgen interrogantes fundamentales en torno a la transparencia de los sistemas, la rendición de cuentas institucional y la posibilidad efectiva de impugnar decisiones automatizadas que inciden directamente sobre el ejercicio de derechos fundamentales. ¿Quién debe responder ante una decisión errónea tomada por una máquina? La actual estructura legal internacional carece de mecanismos claros para atribuir responsabilidad penal o civil a sistemas autónomos por lo que la responsabilidad debería recaer en los diseñadores, operadores y autoridades estatales, quienes deberían garantizar la trazabilidad de las decisiones y asumir las consecuencias legales y éticas de su implementación.

Frente a este escenario, se impone la necesidad de consolidar una gobernanza ética y global de la inteligencia artificial, que articule principios jurídicos vinculantes con marcos técnicos claros y mecanismos institucionales efectivos. Las distintas iniciativas impulsadas por organismos internacionales representan pasos relevantes, aunque aún insuficientes, puesto que aún persiste una brecha entre los principios enunciados y su efectiva implementación. En este sentido, el análisis de los principales instrumentos internacionales sobre gobernanza de la inteligencia artificial revela una convergencia creciente en torno a principios fundamentales como la transparencia, la rendición de cuentas, la no discriminación y la supervisión humana. No obstante, persisten diferencias significativas en su alcance, enfoque y fuerza jurídica. El AI Act de la Unión Europea se distingue por su carácter jurídicamente vinculante y su estructura normativa basada en la clasificación de riesgos. Este reglamento representa el modelo más robusto hasta la fecha, al establecer obligaciones claras, mecanismos de supervisión

institucional y sanciones para garantizar el respeto a los derechos fundamentales, especialmente en sectores sensibles como el empleo, la justicia o la sanidad. En contraste, la Recomendación de la OCDE constituye un marco de orientación no vinculante, pero con gran impacto político y técnico. Ha servido como base para el diseño de políticas públicas en numerosos países y promueve un enfoque pragmático que combina valores éticos con recomendaciones operativas dirigidas tanto al sector público como al privado. La Recomendación de la UNESCO aporta una visión más holística, incorporando dimensiones éticas, culturales y filosóficas que trascienden la lógica regulatoria. Al centrarse en la justicia social, la diversidad y el bienestar de las generaciones futuras, introduce una perspectiva humanista fundamental para pensar la IA más allá del marco técnico-jurídico. Por su parte, el informe del ACNUDH y la Declaración de Toronto ofrecen una aproximación desde los derechos humanos, enfocándose en los riesgos de los sistemas automatizados sobre grupos vulnerables. Ambos destacan la urgencia de garantizar mecanismos efectivos de reparación y atribución de responsabilidades, especialmente frente a fenómenos como la discriminación algorítmica o la opacidad de los sistemas.

En conjunto, estos marcos reflejan una evolución desde guías éticas voluntarias hacia estructuras regulatorias más exigibles, conformando una gobernanza híbrida que combina principios normativos, herramientas jurídicas y marcos técnicos con el objetivo de asegurar un desarrollo justo, seguro y centrado en la dignidad humana de la inteligencia artificial. El futuro de la inteligencia artificial no dependerá exclusivamente de los avances técnicos, sino de la capacidad colectiva para orientar dichos avances hacia fines compatibles con la equidad, la inclusión y la dignidad humana. Sólo bajo estas condiciones será posible que la IA se convierta en una herramienta al servicio del bien común y no en un nuevo instrumento de reproducción de desigualdades estructurales.

Bibliografía.

ALLISON, G., *Destined for War: Can America and China Escape Thucydides's Trap?*, Houghton Mifflin Harcourt, Boston, 2017. ISBN 978-1328915382.

BOLUKBASI, Tolga et al. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, p. 4349–4357.

BUOLAMWINI, J., GEBRU, T., “Gender Shades: Interseccional Accuracy Disparities in Commercial Gender Classification”, *Proceefing of Maching Learning Reserach*, 81, 2018, pp.1-15.

CASTELLANOS, J., “Human Rights, Vulnerability and Artificial Intelligence: An Analysis in Constitutional Perspective”, *Deusto Journal of Human Rights*, nº 14, 2024, pp. 33-50.

CORRALES, José M. “Derecho al trabajo y discriminación: las políticas públicas de empleo como instrumento de igualdad”, en Elena García-Antón (dir.), *Los derechos humanos en la inteligencia artificial: su integración en los ODS de la Agenda 2023*. Cizur Menor: Thomson Reuters Aranzadi, 2022, pp. 399-422.

DASTIN, J., “Amazon scraps secret AI recruiting tool that showed bias against women”, *Reuters*, 10 oct. 2018. Disponible en: <https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/>

FLORIDI, L. et al. “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations”. *Minds and Machines*, vol. 28, 2018, pp. 689–707. doi:10.1007/s11023-018-9482-5.

ICRAC (International Committee for Robot Arms Control), *Stop Killer Robots Campaign Briefing Paper*. 2022. Disponible en: <https://www.stopkillerrobots.org>

LUM, K.; ISAAC, W., “To predict and serve? Predictive policing and the ethics of preemption”, *Significance*, vol. 13, n.º 3, 2016, pp. 14-19.

MOROZOV, E., *Digital Democracy, Analogue Politics: How the Internet Era is Transforming Kenya*. Zed Books, London, 2019. ISBN 978-1786992101

O’NEIL, C., *Armas de destrucción matemática. ¿Cómo el Big Data alimentará la desigualdad y la amenaza democrática?*, Capitán Swing Libros, Madrid, 2018. ISBN: 9788494740848

OBERMEYER, Z.; POWERS, B.; VOGELI, C.; MULLAINATHAN, S. “Dissecting racial bias in an algorithm used to manage the health of populations”. *Science*, vol. 366, n.º 6464, 25 oct. 2019, pp. 447-453. DOI: 10.1126/science.aax2342.

OCDE, *Recomendación del Consejo sobre la inteligencia artificial*, Traducción no oficial, París, 2019. Disponible en: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>

OFICINA DEL ALTO COMISIONADO DE LAS NACIONES UNIDAS PARA LOS DERECHOS HUMANOS, *El derecho a la privacidad en la era digital: Informe de la Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos*. Consejo de Derechos Humanos, 51.º período de sesiones, temas 2 y 3. Documento de la ONU A/HRC/51/17 [en línea]. Ginebra: Naciones Unidas, 4 ago 2022. Disponible en: <https://undocs.org/es/A/HRC/51/17>

PARISER, E., *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press, New York, 2011. ISBN 9781101515129.

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 se modifican determinados actos legislativos de la Unión (Ley de Inteligencia Artificial). Diario Oficial de la Unión Europea, L, 12.07.2024. Disponible en: <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32024R1689>

SHARKEY, N., *The Ethics of Autonomous Weapons Systems*, ne KATINA, P. et al. (eds.), *The Ethics of Artificial Intelligence*. Springer, 2019. ISBN 978-3-030-21834-6.

The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems. 16 de mayo de 2018. Disponible en: <https://www.accessnow.org/the-toronto-declaration/>

UNESCO, *Recomendación sobre la Ética de la Inteligencia Artificial*. París: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, 2021. Disponible en línea: <https://unesdoc.unesco.org/ark:/48223/pf0000381133>.

UNIDIR (United Nations Institute for Disarmament Research). *The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence*. Geneva, 2021. Disponible en: <https://www.unidir.org>

ZHOU, M.; ABHISHEK, V.; KANNAN, S.; et al. *Bias in Generative AI*, University of British Columbia, Universidad de California, Irvine y Carnegie Mellon University, mar. 2024. Disponible en: https://www.andrew.cmu.edu/user/ales/cib/bias_in_gen_ai.pdf

Aitana Torró Calabuig. Los neuro derechos en el horizonte de la singularidad tecnológica

Aitana Torró Calabuig
Universitat de Valencia

[ACCEDE A LA GRABACIÓN](#)

Resumen:

El imparable desarrollo de la inteligencia artificial (IA) y las neurotecnologías ha situado a la mente humana en un escenario de inédita vulnerabilidad, generando un debate jurídico y ético en torno a los llamados neuroderechos. Estas propuestas buscan garantizar la protección de la integridad mental, la privacidad de los datos cerebrales, la libertad cognitiva y la igualdad de acceso a los avances neurotecnológicos. La emergencia de este marco normativo responde al riesgo de que tecnologías capaces de leer, interpretar o incluso manipular procesos neuronales socaven principios fundamentales como la dignidad, la autonomía o la identidad personal. Esta comunicación analiza los principales desafíos que plantea la IA para los derechos humanos, destacando la crisis de identidad derivada de la difuminación de las fronteras entre lo humano y lo tecnológico, así como la necesidad de superar enfoques meramente técnicos (iustecnicismo) mediante un humanismo tecnológico que armonice innovación y valores. En este contexto, los neuroderechos se perfilan como un instrumento innovador para anticipar riesgos y salvaguardar lo más íntimo de la persona: sus pensamientos, recuerdos y emociones. Sin embargo, también se señalan las críticas doctrinales, que advierten sobre el riesgo de redundancia, inflación de derechos o exceso de neuroexcepcionalismo, concluyendo que, más allá de debates conceptuales, resulta imprescindible dotar a los neuroderechos de un contenido jurídico claro, acompañado de evaluaciones de impacto neuroético, marcos regulatorios preventivos y mecanismos de supervisión efectivos. En última instancia, proteger los neuroderechos significa garantizar que, incluso en la era de la inteligencia artificial y la singularidad tecnológica, la mente permanezca como un espacio inviolable y constitutivo de la dignidad humana.

Sumario: 1. Introducción 2. Panorama general de los desafíos éticos de la IA para los derechos humanos 2.1. Crisis de identidad y redefinición de lo humano 2.2. Superar el iustecnicismo: hacia un humanismo tecnológico 3. El caso concreto de los neuroderechos 3.1. Definición y

principios fundamentales 3.2. Iniciativas nacionales e internacionales 3.3. Retos y dilemas jurídicos 3.4. Riesgos de neurodesigualdad y usos indebidos 4. Propuestas para fortalecer el marco jurídico-ético 5. Conclusiones

La mayor causa de alienación en el mundo contemporáneo reside en este desconocimiento de la máquina, que no es una alienación causada por la máquina, sino por el no-conocimiento de su naturaleza y de su esencia, por su ausencia de mundo de significaciones, y por su omisión en la tabla de valores y de conceptos que forman parte de la cultura (Simondon, 2009).

1. Introducción

Un par de meses atrás se podía leer en el periódico una noticia cuyo titular rezaba así: “Un dispositivo lee el pensamiento de personas en tiempo real solo si imaginan la contraseña chittychittybangbang”²⁸². Esta innovadora propuesta científica va encaminada a descodificar frases completas del discurso interior de las personas con parálisis graves que encuentran dificultades o una imposibilidad absoluta para comunicarse mediante el habla, una tecnología conocida como “Brain to Text”. Tras el lógico asombro y entusiasmo que asoman cuando se comienza a leer esta noticia, sigue la incertidumbre y el desasosiego que lleva a realizarse la siguiente pregunta: ¿existe un marco legislativo suficientemente garantista en términos de neuroderechos, de forma que se protejan los datos cerebrales de las personas usuarias?

El Estado chileno lo tuvo claro en 2021 cuando se convirtió en el primer Estado²⁸³ en otorgar tutela constitucional a la actividad cerebral, al afirmar que el desarrollo científico y tecnológico deberá realizarse al servicio de las personas, con respeto a su vida y su integridad física y psíquica, y con un aparato legal que enmarque su uso y sus límites de forma que se resguarde especialmente la actividad cerebral y la información en ella contenida²⁸⁴. Por su lado, en el parlamento cántabro se está tramitando un anteproyecto de ley sobre salud digital que, siendo pionera en Europa en la materia, va encaminada a proteger los neuroderechos y la información cerebral frente a la extracción de datos por parte de dispositivos electrónicos que permiten su comercialización en forma de objetos digitales. De esta forma, y siguiendo la concepción de Yuk Hui (2023), dichos objetos no son simples ficheros informáticos, sino entidades formadas

²⁸² ANSÈDE, Manuel. *Un dispositivo lee el pensamiento de personas en tiempo real solo si imaginan la contraseña “chittychittybangbang”*. EL PAÍS, 14 ag. 2025, 17:01 CEST.

²⁸³ La aprobación en el Estado de Chile produjo una suerte de efecto dominó, adoptándose también legislación similar el estado de Río Grande do Sul, en Brasil, y los de Colorado, California y Montana, en Estados Unidos.

²⁸⁴ Art. 19-1 de la Constitución chilena.

por datos y metadatos cuya existencia se sostiene en una compleja red de relaciones técnicas y sociales, lo que les confiere el estatus de una nueva clase de objeto industrial capaz de circular, acumular valor y reconfigurar las dinámicas económicas y políticas en torno a la información personal más íntima: la actividad cerebral.

En este sentido, el vertiginoso avance de la inteligencia artificial (IA)²⁸⁵ y las tecnologías emergentes plantea desafíos sin precedentes para los derechos humanos y la ética. Se vislumbra en el horizonte la llamada “singularidad tecnológica”, un punto hipotético en que las máquinas podrían igualar o superar la inteligencia humana, revolucionando la relación entre seres humanos y tecnología. Estas intuiciones estuvieron ya muy presentes movimientos culturales y artísticos predecesores de movimientos reaccionarios del S. XX, como es el caso de Marinetti y el fascismo italiano, quien afirmará que “Un automobile ruggente, che sembra correre sulla mitraglia, è più bello della vittoria di Samotracia”, esto es, “un automóvil rugiente, que parece correr sobre la metralla, es más bello que la Victoria de Samotracia”. La fascinación – casi libidinal – por las máquinas, la tecnología o la velocidad desde posiciones neofascistas y profundamente liberales sigue actualmente tan presente como entonces, tal y como refleja la figura de Elon Musk. Así, la desactivación de la potencialidad totalitaria de las tecnologías y de su posible contribución a la continuidad del sistema capitalista pasa necesariamente por el establecimiento de consensos ético-políticos firmes sobre el uso de estos artefactos, que refleje un claro compromiso con los derechos humanos sin paralizar al avance científico en beneficio de toda la humanidad.

En todo caso, volviendo al concepto de singularidad tecnológica²⁸⁶, este fue atribuido en sus primeras acepciones a John von Neumann a mediados del siglo XX y sería ampliamente teorizado posteriormente por autores como Kurzweil, Vernor Vinge o Yampolskiy. Se trata así de una terminología importada de las ciencias puras, puesto que la singularidad alude un punto que tiende hacia el infinito donde impera la incertidumbre. En la filosofía de la tecnología, ha conocido numerosas acepciones: desde la visión de Ray Kurzweil, que la interpreta como el resultado inevitable de una aceleración exponencial del progreso tecnológico hasta la fusión hombre-máquina, hasta la de Vernor Vinge, que la compara con un “horizonte de sucesos” más allá del cual es imposible prever el futuro. Otros enfoques, como los de Anders Sandberg,

²⁸⁵ Autores como Éric Sadin (2020: 35) cuestionarán el uso del término inteligencia artificial como si de una imitación de la inteligencia humana se tratara, dado que los códigos binarios en los que se funda la primera no pueden nunca llegar a alcanzar la aprehensión multisensorial del medio que poseen los humanos, imposible de sistematizar.

²⁸⁶ Se trata de un término muy popularizado actualmente, hasta el punto de existir una universidad a la que da nombre (Universidad de la Singularidad de Silicon Valley).

plantean la existencia de múltiples singularidades parciales, cada una impulsada por distintas tecnologías disruptivas con ritmos de avance propios (Gayozzo, 2021: 197-198).

La propuesta de Kurzweil (2012), uno de los máximos impulsores de esta idea, defiende que la capacidad de los sistemas inteligentes de perfeccionarse a sí mismos acabarán por crear una entidad global inteligente, situando el comienzo de una sociedad postbiológica en 2045, cuando las máquinas se perfeccionarán tanto a sí mismas que todo quedará bajo su control, incluyendo los recursos materiales y energéticos necesarios para mantener su crecimiento, e iniciarán así su expansión cósmica. Mientras que Kurzweil no tiene una percepción pesimista de los escenarios futuros, que abrirán a su parecer nuevos horizontes para el desarrollo de la existencia humana, Vinge sí que sostiene que será el fin de la condición de lo humano. En todo caso, no hay duda de que la humanidad se enfrenta a un reto de una magnitud ingente.

Este punto de inflexión histórico respecto a la relación del ser humano con la naturaleza y con la tecnología ha venido a instaurar un nuevo orden neuro-sanitario, en el que los algoritmos, las nubes de almacenamiento de información y todo tipo de dispositivo digital que hace uso de la IA ha adquirido una relevancia innegable. Ante esta perspectiva, es imprescindible repensar los marcos éticos y jurídicos vigentes, de forma que el Derecho no se quede atrás en el realismo tecnológico en el que se adentara la humanidad. En línea con lo planteado por Llano (2024), una forma de afrontar estos desafíos es mediante el humanismo tecnológico que reconcilie el progreso científico-tecnológico con los valores humanistas fundamentales, suponiendo así una alternativa al conocido como iustecnicismo, una postura dominante que tiende a abordar el Derecho de forma meramente técnica y formalista, desconectada de consideraciones éticas más amplias.

En la presente comunicación se analizarán, de forma no exhaustiva, los principales problemas éticos y jurídicos generados por la IA en relación con los derechos humanos, proponiendo una reflexión crítica sobre la necesidad de reconocer y proteger los neuroderechos como pilar esencial en la era de la singularidad tecnológica. Más allá de los desafíos técnicos, se trata de salvaguardar elementos constitutivos de la dignidad humana, como la integridad moral, la privacidad de los datos cerebrales y la autonomía cognitiva frente a tecnologías capaces de leer, interpretar o incluso modificar la actividad neuronal. El objetivo irá dirigido, por lo tanto, a examinar cómo el Derecho, desde una perspectiva de derechos humanos, puede y debe anticiparse a estos retos para garantizar que el progreso científico y la innovación neurotecnológica se desarrollen en un marco ético y jurídico plenamente respetuoso con la dignidad humana.

Así, se parte de la constatación de que la aceleración del desarrollo de interfaces cerebro-ordenador, unida al avance de la inteligencia artificial, está generando un ecosistema donde la

mente humana se convierte en un espacio potencial de interacción tecnológica directa. Esto plantea, por primera vez en la historia, la necesidad de proteger jurídicamente procesos mentales, no solo frente a injerencias físicas o comunicativas, sino frente a la extracción, análisis y eventual manipulación de la actividad neuronal misma.

2. Panorama general de los principales desafíos éticos de la IA para los DDHH

El impacto de la inteligencia artificial trasciende el ámbito tecnológico o meramente científico para instalarse en el corazón mismo de la reflexión iusfilosófica. De esta forma, más que una mera herramienta, la IA se está convirtiendo en un nuevo actor que reconfigura la manera en que se piensa, se establecen relaciones y se ejercen derechos. Este cambio de paradigma exige ir más allá de la simple adaptación normativa para abordar cuestiones de fondo: ¿qué significa ser humano en un mundo donde las capacidades, hasta ahora exclusivas de la naturaleza humana, pueden ser replicadas o superadas por sistemas artificiales? ¿Cómo garantizar que la regulación tecnológica no se limite a corregir problemas a posteriori, sino que anticipe riesgos y proteja valores fundamentales?

En este marco, los desafíos pueden organizarse en dos planos complementarios: por un lado, la redefinición de lo humano frente a la creciente integración de la tecnología en nuestra vida personal y colectiva; por otro, la superación del iustecnicismo, impulsando un humanismo tecnológico que armonice innovación y derechos.

2.1. Crisis de identidad y redefinición de lo humano

La progresiva integración de la razón tecnológica en la condición ontológica del ser humano suscita profundas inquietudes desde el punto de vista de su regulación. Así, la paulatina autodeterminación de las máquinas dotadas de IA, la expansión de mundos virtuales y la inmersión de las personas en metaversos digitales enfrenta al ser humano a una crisis de identidad, en la que línea divisoria entre lo real-natural y lo digital-artificial se difumina cada vez más, implicando una transformación fenomenológica en la manera en que las personas experimentan, perciben y narran su propia existencia. La integración de prótesis cognitivas, asistentes virtuales y entornos inmersivos no solo amplifica capacidades, sino que también desplaza los límites de la experiencia humana hacia espacios donde la percepción de lo “auténtico” se vuelve ambigua.

En este contexto, uno de los primeros retos que presenta el desarrollo de la inteligencia artificial es el de definir los contornos de lo que se entiende como esencialmente humano, entre otras razones, porque esta es la bisagra que permite conectar a los seres humanos con el consagrado conjunto de derechos humanos. En efecto, conceptos jurídicos tradicionales como sujeto de derechos, dignidad humana, autonomía de la voluntad o intimidad personal deben ser repensados en una era en la que la frontera entre lo humano y lo tecnológico resulta

cada vez más porosa. El conocido filósofo John Searle propondría a través de su famoso experimento sobre la “habitación china” una respuesta al test de Turing, a través del cual tratará de demostrar cómo la lógica usada por las computadoras no alcanza a descifrar el contenido de una acción, sino solo su forma sintáctica o gramatical, al contrario que la humana, que tiene capacidad semántica para entender el significado de los símbolos. De esta forma, Searle (1984: 28) rechazará la equiparación entre cerebro/hardware y mente humana/software, en tanto que la computación se fundamenta en un lenguaje binario que carece de intencionalidad y de semántica.

Asimismo, con el objeto de ofrecer una respuesta convincente a esta pregunta, nacen corrientes del pensamiento como las del posthumanismo o transhumanismo. En 1990, Max Moore definiría el transhumanismo como un conjunto de filosofías que buscan guiar hacia una condición poshumana, compartiendo muchos elementos con el humanismo (el respeto por la razón y la ciencia, un compromiso con el progreso y una apreciación de la existencia humana (o transhumana) antes de la muerte. No obstante, difiere del humanismo al reconocer anticipar las alteraciones radicales en la naturaleza y en las posibilidades vitales que resultarán del desarrollo de diversas ciencias y tecnologías, como la neurociencia y la farmacología, las investigaciones sobre la extensión de la vida, la nanotecnología, la ultrainteligencia artificial, la exploración del espacio, combinado todo ello con una filosofía y un sistema de valores racionales.

Así, dos avances científicos produjeron un cambio sustancial en la forma de aproximarse a lo humano: la cibernética y la biotecnología (Chavarría, 2015). La cibernética introdujo a mediados del S. XX un paradigma centrado en la teoría de sistemas, el control y la información, desplazando al sujeto racional como centro explicativo, como ilustra la teoría del actor-red de Bruno Latour. Así, el ser humano comenzó a ser concebido como un procesador de información, análogo a las máquinas inteligentes, lo que condujo a la descorporeización de la razón (pues el razonamiento se desanclaría del cuerpo humano) y a la disolución de las fronteras entre lo humano y lo no humano.

Más adelante, las biotecnologías (genética, biología molecular, técnicas reproductivas) reforzaron la somatización de la identidad, que pasó a entenderse en términos moleculares y neuroquímicos. Así, la corporeidad se ha ido convirtiendo progresivamente en objeto de intervención y mejora (mediante fármacos, cirugía, manipulación genética), mientras se consolidaba un discurso social que asocia la existencia humana al plano corporal y a procesos metabólicos, un desplazamiento que implica riesgos de biologización de la cultura, determinismo genético, control biopolítico y subordinación a la autoridad científico-médica (Rose, 2007).

Será así como el posthumanismo se nutrirá de estos cambios, proponiendo superar el humanismo antropocéntrico a través de un mundo postsocial, en el que humanos, máquinas e inteligencias artificiales coexisten en horizontalidad. Sus representantes, como Pepperell, Moravec o Minsky, y el movimiento transhumanista, defienden intervenir en la naturaleza humana para trascender los límites biológicos y alcanzar estados posthumanos: sujetos autoprogramados, potencialmente inmortales, sin esencia espiritual ni centralidad ontológica. También Sloterdijk, desde una perspectiva filosófica, aboga por nuevas antropotécnicas que continúan la tradición de autotransformación humana mediante técnicas y biotecnologías.

En la medida en que los sistemas de IA pueden emular, y en algunos casos superar, habilidades humanas como la creatividad, el razonamiento lógico o la interacción lingüística, surge la necesidad de preguntarse si tales capacidades son meramente imitaciones funcionales o si constituyen un equivalente sustantivo de la cognición humana. Esta discusión, además de filosófica, tiene profundas implicaciones jurídicas: si la IA llega a ser considerada un agente con algún grado de autonomía reconocible, el Derecho deberá replantearse las categorías de responsabilidad, agencia y capacidad jurídica.

La pregunta de la naturaleza humana, cuya respuesta excede los límites del presente texto, debe reconducirnos a pensar en la necesidad de definir marcos normativos innovadores que estén a la altura del doble desafío actual: evitar que las tecnologías erosionen derechos ya reconocidos —como la privacidad, la libertad de pensamiento o la igualdad—; por otro, identificar y consagrar nuevos derechos emergentes, como los neuroderechos, capaces de responder a los riesgos inéditos que surgen cuando las máquinas acceden, procesan o manipulan información directamente derivada de la actividad cerebral. De este modo, la identidad humana deja de ser un ámbito exclusivamente propio y se convierte en un terreno susceptible de intervención tecnológica.

Para ello, es importante desembarazarse del formalismo tecno-jurídico o iustecnicismo, un enfoque que tiende a concebir la regulación de la IA y las tecnologías digitales como un problema meramente técnico, abordable mediante ajustes normativos aislados, sin cuestionar los fundamentos éticos o filosóficos del Derecho. Asimismo, la IA aplicada a entornos sensoriales y cognitivos —como la visión artificial, el procesamiento del lenguaje natural o la lectura cerebral— puede erosionar barreras históricas entre el “espacio privado interior” y el “espacio público exterior”, obligando a replantear el alcance de la intimidad, la libertad de pensamiento y el derecho a no ser manipulado cognitivamente.

2.2. Superar el iustecnicismo: hacia un humanismo tecnológico

El humanismo tecnológico constituye una reformulación del humanismo en el contexto de los actuales desarrollos científico-tecnológicos, en especial de la inteligencia artificial. Su propósito es superar la dicotomía entre ciencias y humanidades —la llamada teoría de las “dos culturas”— mediante una aproximación integral e interdisciplinaria que reconozca la complejidad del ser humano en todas sus dimensiones. Entre otras cosas, el humanismo tecnológico se opondrá a la idea de una esencia humana inmutable y, a diferencia de las visiones tecnocráticas de corte determinista o de los discursos que plantean una supuesta deshumanización de la técnica, el humanismo tecnológico concibe la tecnología como un horizonte de posibilidades cuya orientación debe estar guiada por criterios éticos de responsabilidad y sostenibilidad (Terrones, 2019). No se trata de recrear paradigmas humanistas pasados, sino de articular un marco filosófico capaz de contextualizar al ser humano en el presente “tiempo tecnológico”, fomentando una educación integral, una cultura de corresponsabilidad y una praxis orientada a un desarrollo tecnológico responsable y respetuoso de la dignidad humana.

Ante este nuevo paradigma jurídico, en el sentido de Kuhn, resulta fundamental ofrecer una formación más holística que rompa con los silos disciplinares. Así, no solo es necesario que existan personas juristas familiarizadas con conceptos propios de la ciencia de datos, los algoritmos o la ética de la IA, sino también personas científicas e ingenieras que tengan nociones básicas de derechos humanos y principios jurídicos. Así, mediante este diálogo transversal se pueden formular soluciones innovadoras que sean técnica y éticamente sólidas, lo cual implica incorporar la formación en ética y derechos humanos en los programas educativos tanto de las carreras técnicas como de las jurídicas. Las carencias en formación ética y técnica en ambas direcciones pueden llevar a que los desarrolladores de IA ignoren las implicaciones sociales de sus creaciones, o que, paralelamente, los legisladores no comprendan cabalmente cómo funcionan los sistemas algorítmicos que intentan regular.

La intersección entre tecnología y derecho ya no puede resolverse con “parches” normativos que persigan a posteriori los avances científicos; requiere marcos regulatorios anticipatorios, evaluaciones de impacto ético obligatorias y la incorporación de principios de precaución tecnológica. Esto implica también aceptar que la regulación debe ser dinámica, revisable y flexible, pero siempre asentada sobre valores humanistas que sirvan de brújula frente a la volatilidad del cambio tecnológico.

En síntesis, superar el iustecnicismo implica humanizar la tecnología y tecnificar al humanista en igual medida. Este equilibrio requiere abandonar tanto el determinismo tecnológico —que asume que todo avance es inevitable e incontrolable— como el inmovilismo jurídico —que

desconfía de la innovación y busca frenarla por defecto—. La clave está en promover un humanismo tecnológico que oriente la investigación y el desarrollo hacia el bien común, incorporando criterios de justicia distributiva, sostenibilidad y respeto a la diversidad cultural y cognitiva.

Se trata de equilibrar la balanza: ni un tecnocentrismo ciego a los valores, ni un romanticismo jurídico alérgico a la innovación. El humanismo tecnológico aboga por un progreso con conciencia, donde cada avance en IA sea evaluado a la luz de su contribución al bienestar humano y su compatibilidad con los derechos fundamentales. Bajo esta perspectiva, la pregunta rectora ya no es solo qué puede la tecnología hacer, sino también qué debe hacer y bajo qué condiciones. Y la respuesta a esta última debe construirse participativamente, incorporando voces de filósofos, científicos, juristas, sociólogos y de la ciudadanía en general. Solo así podremos trazar una hoja de ruta ética para la IA que evite sus peligros y potencie sus beneficios en favor de la sociedad.

3. El caso concreto de los neuroderechos

El debate sobre los neuroderechos se despliega así en múltiples frentes: desde la necesidad de precisar su definición y principios fundamentales, hasta la clasificación de los distintos derechos que se proponen bajo esta etiqueta; desde las iniciativas legislativas pioneras en Chile y los foros multilaterales que han comenzado a incorporarlos en sus agendas, hasta las críticas doctrinales que advierten del riesgo de inflacionar el catálogo de los derechos humanos. Todo ello se complementa con los dilemas jurídicos derivados de la protección de los neurodatos, el acceso equitativo a mejoras cognitivas o la eventual militarización de las neurotecnologías. A la luz de esta complejidad, el primer paso consiste en examinar qué entendemos por neuroderechos y cuáles son los principios sobre los que se sustentan.

3.1. Definición y principios fundamentales

Entre las propuestas más vanguardistas en el cruce de tecnología y derechos humanos destaca la de los neuroderechos, un término que alude a una nueva generación de derechos fundamentales orientados a proteger el cerebro, la mente y la identidad personal (en definitiva, la psique en su integridad) frente a potenciales amenazas derivadas de las neurotecnologías. El núcleo que este tipo de derechos pretende salvaguardar es la integridad mental, la libertad cognitiva (libre albedrío y autodeterminación sobre los propios procesos neuronales), la privacidad mental (protección de los datos e información cerebral) y la igualdad de acceso a mejoras neurotecnológicas para evitar una brecha neuro-tecnológica entre personas. En este sentido, Mascetti (2022) afirmará que “el homo sapiens se distingue por sus actividades cognitivas. Por ello, los neuroderechos son derechos humanos, ya que se refieren particularmente a ellas y, como tales, son universales”.

En todo caso, el debate sobre los neuroderechos se articula en torno a la idea de que el desarrollo de las neurotecnologías genera riesgos inéditos para la autonomía, la privacidad y la identidad humana, que exigen respuestas jurídicas específicas. Desde una primera perspectiva, parte de la literatura científica sostiene que deberían reconocerse al menos cuatro grandes ejes: la privacidad de los pensamientos, la identidad personal, el acceso equitativo a mejoras cognitivas y la protección frente a sesgos en el uso de estas tecnologías. En esta línea, la Neurorights Foundation – dirigida por Rafael Yuste – ha sistematizado cinco derechos fundamentales: privacidad mental, identidad personal, libre albedrío, acceso equitativo al aumento cognitivo y protección contra el sesgo algorítmico. La finalidad común de estos derechos sería garantizar que la mente permanezca como un espacio inviolable, preservando la dignidad y el núcleo esencial de la persona en la era digital.

También debe destacarse la propuesta de Ienca y Andorno (2021), quienes distinguen entre libertad cognitiva, privacidad mental, integridad mental y continuidad psicológica. Estas nociones enfatizan que la mente – antes un espacio libre de injerencias – debe ser protegida actualmente de manipulaciones externas y que la percepción de la propia identidad no puede ser erosionada por el uso de tecnologías invasivas. Como apuntaría Boire (2001: 88) en relación con la importancia de la libertad cognitiva, “el derecho a controlar la propia conciencia es la quintaesencia de la libertad. Si la libertad significa algo, debe significar que cada persona tiene el derecho inviolable de pensar por sí misma”, esto es, dirigir su propia conciencia; los propios procesos mentales subyacentes y las propias creencias, opiniones y visión del mundo. Por lo tanto, libertad cognitiva y libertad de pensamiento en la era tecnológica se encuentran intrínsecamente ligados.

En contraste, otros autores han cuestionado la necesidad de crear un catálogo autónomo de neuroderechos, argumentando que muchas de estas garantías ya estarían implícitamente recogidas en derechos clásicos como la intimidad, la integridad física y psíquica o la libertad de pensamiento. Desde esta postura crítica, hablar de neuroderechos como un bloque diferenciado podría ser redundante y generar confusión normativa. No obstante, incluso entre quienes defienden esta tesis, suele reconocerse que la singularidad de los datos neuronales y las posibilidades de manipulación directa de la conciencia justifican, como mínimo, un esfuerzo de reinterpretación de los derechos existentes, cuando no una nueva formulación adaptada a los desafíos específicos que plantean las neurotecnologías.

3.2. Iniciativas nacionales e internacionales

A título de ejemplo, como se ha destacado en la introducción, la propuesta chilena –pionera mundial en este campo– concibe los neuroderechos como “derechos del cerebro” y busca proteger la integridad mental, el libre albedrío y la no discriminación en el acceso a las neurotecnologías. Además, establece que los datos neuronales de una persona gocen de un

estatus especial equiparable al de un órgano, de modo que no puedan ser objeto de tráfico o manipulación indebida. Así, se pretende poner freno a la potencial mercantilización de la información extraída del cerebro, puesto que esta forma parte inherente de la persona humana, no pudiendo ser tratada como un dato comercial más.

A nivel internacional, el concepto de neuroderechos ha ido ganando terreno en la agenda de políticas públicas y foros multilaterales. En 2019, la OCDE aprobó una Recomendación sobre Innovación Responsable en Neurotecnología, primer instrumento intergubernamental que aborda principios para el desarrollo ético de estas tecnologías. Por su lado, el primer pronunciamiento de la ONU sobre los neuroderechos fue el informe *Ethical Issues of Neurotechnology*, elaborado por la UNESCO en 2021. Este documento aborda las interacciones entre neurotecnología, ética y derecho, y plantea recomendaciones para el diseño de marcos de gobernanza adecuados. Posteriormente, el Consejo de Derechos Humanos de Naciones Unidas dio un paso más con la resolución A/HRC/RES/51/3, adoptada el 6 de octubre de 2022, en la que se solicita la elaboración de un estudio específico sobre las implicaciones que la neurotecnología puede tener para la protección y garantía de los derechos humanos.

En el ámbito regional, en 2021 la Organización de Estados Americanos (OEA) expidió la Declaración del Comité Jurídico Interamericano sobre Neurociencia, Neurotecnologías y Derechos Humanos: Nuevos Desafíos Jurídicos para las Américas, que contiene una serie de recomendaciones para los Estados. buscando guías comunes para sus países miembros, el sector privado, la academia y el mundo científico. Más recientemente (octubre de 2023), los ministros de la Unión Europea suscribieron la llamada Declaración de León, primera declaración europea reconociendo la importancia de los neuroderechos y la necesidad de cooperación para su protección.

A escala nacional, el concepto de neuroderechos se encuentra indisolublemente ligado al científico Rafael Yuste, Morningside Group y el proyecto BRAIN (Brain Research Through Advancing Innovative Neurotechnologies), que han sistematizado cinco grandes ejes para su protección: (i) derecho a la identidad personal; (ii) derecho al libre albedrío o libertad cognitiva; (iii) derecho a la privacidad mental; (iv) derecho al acceso equitativo a mejoras neurotecnológicas; y (v) derecho a la protección contra sesgos algorítmicos en el procesamiento de datos neuronales. Estos principios coinciden en otorgar a la información cerebral una categoría especial dentro de la legislación sobre protección de datos, reconociendo su carácter irremplazable y no transferible. De hecho, las acciones del mencionado grupo investigador llegaron a la ONU, en donde el equipo de Yuste abogó para que los neuroderechos fuesen reconocidos como parte del catálogo de derechos humanos incluidos en la Declaración Universal de 1948 (Arellano, 2024: 6).

Cabe destacar que en 2021 se adoptaría la conocida como “Carta de los Derechos Digitales” con el propósito de reforzar los derechos fundamentales proclamados a nivel constitucional, teniendo presente como eje inspirador la garantía de la dignidad humana (art. 1.3). Así, la idea no es tanto crear nuevos derechos sino adaptar los ya existentes al entorno digital, dinámico e imprevisible, aunque esta carta no tendrá carácter vinculante.

3.3 Retos y dilemas jurídicos

No obstante, el reconocimiento formal de los neuroderechos no está exento de desafíos. Uno de los más relevantes es la delimitación jurídica del concepto de “dato neuronal”: ¿incluye solo la actividad eléctrica cerebral en bruto, también los patrones ya decodificados o incluso las inferencias cognitivas que un sistema de IA pueda extraer de ellos? A ello se suma el reto de probar la manipulación o acceso no autorizado en un contexto donde la conversión de impulsos cerebrales a información digital puede desdibujar los límites entre lo estrictamente neuronal y lo puramente informático. Además, existe el riesgo de que, en aras de la seguridad pública o la investigación criminal, se pretenda justificar un acceso coercitivo a estos datos, abriendo una peligrosa grieta en la privacidad mental.

Por otro lado, es importante señalar la amenaza de la neurodesigualdad: la posibilidad de que solo ciertos grupos socioeconómicos puedan acceder a mejoras cognitivas, creando una brecha estructural entre “aumentados” y “no aumentados”. En paralelo, la ausencia de prohibiciones claras sobre usos militares o de control cognitivo masivo podría desembocar en un escenario de militarización de la neurotecnología.

No obstante, la consagración de una nueva generación de derechos humanos vinculada a la neurotecnología también ha generado resistencia. Así, Borbón et. al (2020) apuntan hacia el carácter innecesario de la aprobación de la incorporación de los neuroderechos, por la falta de consenso en cuanto al propio término o lo problemático de conceptos como “el neuroderecho al libre albedrío y al acceso equitativo a neurotecnologías de mejora”, por la fuerte carga ideológica del primero o la falta de deliberación previa respecto a la idoneidad de las tecnologías de mejora y la posible limitación a las que se dediquen en exclusiva a fines terapéuticos.

Como recuerda Asís (2022: 62), en “Neurorights in Chile: Between neuroscience and legal science”, cuatro académicos chilenos realizan un análisis crítico de las iniciativas legislativas presentadas en su país para incorporar los denominados neuroderechos. Según su valoración, estas propuestas carecen de un sustento jurídico sólido y se apoyan en una concepción filosófica que califican como cartesiana y reduccionista, hoy superada, que parte de la premisa de que el cerebro, como órgano singular, requiere un conjunto de derechos exclusivos para su protección. A juicio de los autores, este enfoque resulta problemático porque reproduce

derechos que ya se encuentran amparados en la mayoría de sistemas jurídicos occidentales, como la intimidad, la integridad física y la integridad psíquica. De esta forma, en lugar de formular nuevos derechos con riesgo de solaparse o generar redundancias, consideran más adecuado reforzar la efectividad de los instrumentos jurídicos ya existentes y adaptarlos a los retos específicos que plantean las neurotecnologías.

El neuroeticista Bublitz (2022) también ha sido crítico con la propuesta de los neuroderechos, señalando que la proliferación de nuevas categorías jurídicas genera una inflación que, lejos de fortalecer la protección de los derechos fundamentales, puede terminar devaluándolos. Así, no solo criticará la ausencia de un debate académico riguroso previo que justifique la insuficiencia de los marcos normativos actuales, sino que además la iniciativa está marcada más por un impulso político que por una reflexión jurídica fundada, lo que conlleva una intromisión en terrenos que deberían ser analizados en profundidad por la doctrina y la ciencia jurídica. Además, este autor se mostrará en desacuerdo con lo que él denomina “neuroexcepcionalismo” y “neuroesencialismo”: la tendencia a sobredimensionar las explicaciones de corte neuronal como si fuesen intrínsecamente superiores o exclusivas para comprender la conducta humana. Desde esta óptica, el entusiasmo legislativo en torno a los neuroderechos se explica más por la fascinación hacia el prefijo “neuro” que por una necesidad real.

4. Propuestas para fortalecer el marco jurídico-ético

El Comité Internacional de Bioética de la UNESCO, en su informe de 15 de diciembre de 2021 sobre las implicaciones éticas de la neurotecnología, identifica diversas vías para el reconocimiento y protección de los neuroderechos. Entre ellas, propone la incorporación de protocolos específicos a instrumentos internacionales ya vigentes, como la Declaración Universal de Derechos Humanos, con el fin de responder a los retos que plantea la neurotecnología; el fortalecimiento del propio texto de la Declaración, considerando que las capacidades actuales de estas tecnologías pueden poner en riesgo derechos humanos ya consolidados y que, por ello, sería necesario añadir garantías adicionales; o, de forma más ambiciosa, la redacción de una nueva Declaración Universal de Derechos Humanos y Neurotecnología.

En esta última opción, el Comité plantea que el instrumento normativo debería inspirarse en principios como:

- El reconocimiento del derecho universal a la protección de la actividad cerebral de toda persona, con independencia de su origen étnico, género, situación socioeconómica o capacidades cognitivas.
- La prohibición de utilizar datos cerebrales obtenidos mediante neurotecnología para fines de vigilancia o elaboración de perfiles sin un consentimiento libre e informado, así

como la proscripción de cualquier forma de discriminación basada en características cognitivas o mentales.

- La necesidad de que los usos estatales o privados de la neurotecnología sean evaluados de forma sistemática para prevenir vulneraciones de derechos humanos.
- La importancia de fomentar la información pública, la educación y el debate social sobre la neurotecnología, como vía para garantizar un desarrollo y utilización responsables y éticos de estas herramientas.

En este sentido, resulta imprescindible acompañar el reconocimiento de los neuroderechos con un marco regulatorio robusto y preventivo. Algunas medidas concretas podrían incluir:

- Su **constitucionalización** o incorporación en leyes orgánicas con rango de protección reforzada.
- La **clasificación especial de los datos neuronales** dentro de la normativa de protección de datos, estableciendo un consentimiento informado, libre, específico y revocable, y prohibiendo su cesión comercial.
- La obligación de realizar **evaluaciones de impacto neuroético** antes de la autorización de cualquier interfaz cerebro-máquina o sistema de IA que procese información cerebral.
- La creación de **autoridades independientes de supervisión** que auditen periódicamente tanto la tecnología como el uso de los datos neuronales.
- La negociación de **tratados internacionales** que limiten o prohíban expresamente el uso militar y coercitivo de las neurotecnologías.
- El establecimiento de **mecanismos públicos de acceso equitativo** a tecnologías de mejora o asistencia neurocognitiva para evitar la concentración de capacidades en manos de una élite.

En síntesis, los neuroderechos representan un campo emergente que busca asegurar que el último reducto de libertad humana –la mente– permanezca inviolable. Son una respuesta jurídica a la pregunta de cómo conservar la dignidad y autonomía mental en la era de las interfaces neuronales y la IA ubicua. Si el siglo XX fue testigo de la expansión de los derechos humanos a nuevas esferas (económicas, culturales, ambientales), el siglo XXI podría ver nacer estos derechos neuro-cognitivos. La clave será dotarlos de contenido práctico y armonizarlos con el acervo existente de derechos fundamentales, de forma que no creen contradicciones sino sinergias. Al final del día, proteger los neuroderechos es proteger la esencia de lo que nos hace humanos: nuestros pensamientos, memorias, decisiones y sentimientos más íntimos. Y ello enlaza directamente con el valor de la dignidad humana, reconociendo que incluso ante tecnologías capaces de adentrarse en nuestro cerebro, existe un núcleo inviolable de la persona que ninguna máquina (ni Estado ni corporación) debe traspasar sin consentimiento y garantías.

5. Conclusiones

El debate en torno a los neuroderechos refleja uno de los mayores desafíos jurídicos y éticos de nuestro tiempo: cómo garantizar la dignidad humana frente a tecnologías capaces de acceder, interpretar o incluso modificar la actividad cerebral. La aceleración de la inteligencia artificial y las neurotecnologías sitúa a la mente humana en el centro de un nuevo escenario de vulnerabilidades, obligando al Derecho a repensar categorías clásicas como la intimidad, la autonomía o la identidad personal.

Aunque existen posturas divergentes sobre la necesidad de reconocer neuroderechos como una categoría autónoma —algunos autores defienden reforzar los marcos normativos ya existentes, mientras otros impulsan su formulación explícita—, existe consenso en que los datos neuronales poseen una singularidad que exige protección especial. La privacidad mental, la libertad cognitiva, la identidad personal y el acceso equitativo a los avances neurotecnológicos aparecen como núcleos comunes en las distintas propuestas.

El reto principal es, por tanto, traducir estas ideas en un marco normativo sólido, anticipatorio y global, que evite tanto la mercantilización de la información cerebral como la creación de brechas neurotecnológicas o su utilización con fines de control social y militar. En este sentido, iniciativas como la pionera reforma constitucional chilena, la Declaración de la OEA, los pronunciamientos de la UNESCO o la reciente Declaración de León en Europa, evidencian una creciente conciencia internacional. Sin embargo, más allá de declaraciones de intenciones, resulta indispensable acompañar estos avances con garantías jurídicas efectivas, evaluaciones de impacto neuroético y mecanismos de supervisión independientes.

En definitiva, los neuroderechos constituyen una respuesta emergente ante la singularidad tecnológica. Su reconocimiento no debe entenderse como una moda jurídica, sino como una apuesta por preservar lo más íntimo del ser humano en la era de la inteligencia artificial: el derecho a mantener inviolable la mente. La clave será articularlos en armonía con el corpus de derechos humanos ya existente, asegurando que el progreso tecnológico se desarrolle bajo principios de justicia, igualdad, libertad y respeto a la dignidad humana.

Referencias

ARELLANO TOLEDO, W. Los Neuroderechos y su regulación. *Inteligencia Artificial*, 2024, vol. 27, nº 73, pp. 4-13. <https://doi.org/10.4114/intartif.vol27iss73pp3-1>.

ASÍS, Rafael de. Sobre la propuesta de los neuroderechos. *Derechos y Libertades*, 2022, nº 47, pp. 51-70. <https://doi.org/10.20318/dyl.2022.6873>

BOIRE, Ricard Glen. On cognitive liberty III. *Journal of Cognitive Liberties*, 2001, vol. 2, nº 1.

BORBÓN RODRÍGUEZ, Diego Alejandro; BORBÓN RODRÍGUEZ, Luisa Fernanda; LAVERDE PINZÓN, Jeniffer. Análisis crítico de los NeuroDerechos Humanos al libre albedrío y al acceso equitativo a tecnologías de mejora. *Ius et Scientia*, Sevilla, Universidad de Sevilla, 2020, vol. 6, nº 2, pp. 135–161. <https://doi.org/10.12795/IETSCIENTIA.2020.i02.10>.

BUBLITZ, Jan Christoph. (2022). Novel Neurorights: From Nonsense to Substance, *Neuroethics*, 2022, vol. 15, nº 7. <https://doi.org/10.1007/s12152-022-09481-3>

CHAVARRIA ALFARO, Gabriela. El posthumanismo y los cambios en la identidad humana. *Reflexiones*, 2015, vol.94, nº 1, pp. 97-107. http://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S1659-28592015000100097&lng=en&nrm=iso.

GAYOZZO HUAMANCHUMO, Piero Arturo. Singularidad tecnológica y transhumanismo. *Teknokultura. Revista de Cultura Digital y Movimientos Sociales*, 2021, vol. 18, nº 2, pp. 195-200. <http://dx.doi.org/10.5209/TEKN.74056>.

HUI, Yuk. Sobre la existencia de los objetos digitales. Segovia: *Materia Oscura*, 2023. Traducción de José Luis de la Nuez Santana.

IENCA Marcello Y ANDORNO, Roberto. Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci Soc Policy*, 2017, 13(5), pp. 141-185.

KURZWEIL, Raymond. La singularidad está cerca. Cuando los humanos trascendamos la biología. Berlín, Lola books, 2012.

LLANO ALONSO, Fernando H. *Homo Ex Machina. Ética de la inteligencia artificial y Derecho digital ante el horizonte de la singularidad tecnológica*. Valencia: Tirant lo Blanch, 2024.

MASCITTI, M. El rango constitucional de los neuroderechos como una exigencia de justicia. *Cuestiones Constitucionales*, 46: 149- 176, 2022.

MORE, Max. Transhumanism: Toward a Futurist Philosophy, *Extropy*, 1990, nº 6, pp. 6-11.

NEURORIGHTS FOUNDATION. The Five NeuroRights. Disponible en: <https://www.neurorightsfoundation.org/mission/impact>.

ROSE, Nikolas. *The Politics of Life Itself, Biomedicine, Power, and Subjectivity in the Twenty-First Century*, 2007, Princenton University Press.

SADIN, Éric (2020). *La inteligencia artificial o el desafío del siglo*. Buenos Aires: Caja Negra Editora.

SEARLE, John R. (1984). *Minds, Brains and Science*. Cambridge: Harvard University Press.

SIMONDON, Gilbert. *La individuación a la luz de las nociones de forma y de información*, 2009, Buenos Aires: Cactus-La Cebra.

TERRONES, Antonio Luis. Humanismo tecnológico. Fundamento para una inteligencia artificial responsable. *Pensamiento Actual*, vol. 19, nº. 33, 2019, pp. 15-24.

UNESCO, Recomendación sobre la Ética de la Inteligencia Artificial, 41a Conferencia General, 2021. Unión Europea, Propuesta de Reglamento de IA (Artificial Intelligence Act), COM(2021) 206 final. Tribunal Supremo de Chile, Sala Constitucional, Sentencia Rol 105065-2023 (caso neuroderechos/ Emotiv)

Carmen Ventura Salom. Retos y desafíos que plantea la Inteligencia artificial en las actuales democracias

Carmen Ventura Salom.

Doctora por la Facultad de Derecho de la Universitat de València.
Máster en Estudios Internacionales y de la Unión Europea. Licenciada en CC. Políticas y de la Administración.

[ACCEDE A LA GRABACIÓN](#)

Resumen

El objeto de estudio de la comunicación es analizar la importancia del respeto a los derechos fundamentales y a las libertades públicas, que subyacen en las democracias actuales, en la aplicación de sistemas de inteligencia artificial en ámbitos considerados como esenciales en la estructura social, económica y política de la ciudadanía.

Existe una gran cantidad de estudios que desde multiplicidad de variables plantean y debaten sobre la notable influencia de las nuevas tecnologías y particularmente de la revolución que va a suponer la normalización del uso de la inteligencia artificial (en adelante IA) en la vida diaria de las personas.

Los avances desarrollados a partir de las teorías del matemático británico Alan Turing, considerado como el padre de la inteligencia artificial, han supuesto, como afirma Oliver, “una revolución en la investigación y creación de sistemas que realizan tareas propias de la inteligencia humana como la percepción, el razonamiento y el aprendizaje ya que asistimos al hecho de que programas de ordenador puedan crear contenido textual (y también imágenes, música y vídeos) indistinguible del que generaría un humano pero con una escala y velocidad muy superiores a los de los humanos²⁸⁷

²⁸⁷Oliver N., *¿La pluma inteligente?*, consultado en la página web de la Real Academia de Ingeniería: <https://www.raing.es/la-pluma-inteligente-por-nuria-oliver/> (artículo publicado en el diario ABC el 14 de abril de 2023)

Esta nueva incorporación de algoritmos en las distintas áreas de actuación de la ciudadanía supondrá, sin duda, una nueva dinámica relacional en la que consideramos deberá prevalecer la razón, la ética y las buenas praxis ya que de lo contrario derechos tales como la privacidad o el derecho a la imagen podrían verse afectados y especialmente entre los colectivos más vulnerables.

En este contexto de transformación y adaptación, reflexionaremos sobre las oportunidades, los retos y los riesgos que suponen el desarrollo y la aplicación del uso masivo de datos y su impacto en los distintos ámbitos relacionales de las personas, concretamente en aquellas esferas en las que deben prevalecer la protección a la integridad e intimidad de la ciudadanía, evitando así la explotación y utilización de datos sensibles que podrían suponer una discriminación y por tanto una vulneración de los derechos humanos y de las libertades fundamentales reconocidos en la Declaración Universal de los Derechos Humanos²⁸⁸.

Abordar una cuestión tan compleja, novedosa, diferente e imprevisible requiere de un marco de actuación transversal que vincule a las administraciones e instituciones públicas, a las empresas y a la sociedad civil, en base a unos criterios que garanticen, protejan y afiancen los derechos fundamentales en entornos digitales ya que como afirma Cotino: “todos los sistemas automatizados o algorítmicos potencialmente pueden generar impactos y afectaciones a los derechos de las personas u otros bienes e intereses protegibles”²⁸⁹

Consideramos, por tanto, que el desarrollo y la utilización de la IA es una responsabilidad colectiva y compartida por todos los miembros de una sociedad y su finalidad, desde una perspectiva ética y respetuosa con los derechos humanos y libertades fundamentales, debe contribuir al progreso y al bienestar del conjunto de la ciudadanía, de lo contrario, podríamos adentrarnos en un escenario en el que más allá de su complejidad, las personas se vean despojadas de sus derechos esenciales y pasen a convertirse en meros receptores pasivos que bien sea por razones ideológicas o económicas puedan ver limitadas o

²⁸⁸Declaración Universal de los Derechos Humanos. Consultado en la página web de las Naciones Unidas: <https://www.un.org/es/about-us/universal-declaration-of-human-rights>.

²⁸⁹Cotino, L. *Implantando la transparencia y los registros públicos de algoritmos. La experiencia pionera en la Comunidad Valenciana*. *Revista Valenciana d'Estudis Autònoms*, 2023, nº 68,p. 61.

reducidas sus posibilidades de actuación en sociedades fragmentadas, individualistas, carentes de empatía y de solidaridad.

1. Introducción

La reciente incorporación de la IA en los distintos ámbitos en los que las personas interactúan y sus consecuencias, han puesto de relieve la necesidad de establecer unos nuevos cimientos que respondan a las características propias de un escenario diferente, complejo e incierto que se caracteriza por la aparición de nuevas dinámicas relacionales tanto sociales como políticas y económicas.

Conviene tener en cuenta que la irrupción de la IA acontece en un escenario geopolítico de contiendas bélicas, crisis humanitarias y económicas y una fuerte polarización social.

En este sentido, el diseño y desarrollo de los sistemas de IA responde a diferentes escenarios internacionales, en el que distintas potencias: China, EEUU y los 27 Estados miembros de la UE se disputan la hegemonía.

En su análisis, Ortega establece diferencias entre el “sistema estadounidense, o “capitalismo de vigilancia”, en el que un pequeño número de proveedores de servicios digitales disfruta de un poder de mercado preponderante y de ventajas informativas digitales, especialmente de control, sobre los datos de los usuarios y sobre la publicidad.

El chino, o “Estado de vigilancia”, en el que el Estado goza de ventajas informativas preponderantes, apoyado por un pequeño número de proveedores de servicios digitales. En el que Estado quiere centralizar los datos, porque son un instrumento de control político, y

El europeo, que busca una regulación centrada en las personas, o “humanista” y en el que los proveedores de servicios digitales se ven constreñidos por una regulación de la UE que se centra en la protección de los valores y derechos de los usuarios digitales²⁹⁰.

En este contexto globalizado y asimétrico, en opinión de Lasalle “el caso chino está claramente vinculado a un paradigma de Estado-plataforma que ejerza un control absoluto sobre su población y que dote a China de un poder hegemónico y en Estados Unidos es una propuesta neoliberal que traslada esa capacidad hegemónica a las corporaciones, que

290 Ortega, A. *Hacia un régimen europeo de control de la Inteligencia Artificial*, consultado en página web del Real Instituto Elcano: <https://www.realinstitutoelcano.org/analisis/hacia-un-regimen-europeo-de-control-de-la-inteligencia-artificial/>

desarrollan un mercado digital basado en el egoísmo, los deseos, los anhelos más inconscientes que definen al ser humano cuando se relaciona con la tecnología”²⁹¹

En esta reciente dinámica relacional de enfrentamiento e intolerancia entre los ciudadanos y dadas las tensiones sociales y políticas que han polarizado ideológicamente a gran parte de las sociedades democráticas occidentales, De Bartolomé plantea la “necesidad de estructurar, adecuar y desarrollar, una normativa que regule tanto los contenidos de los sistemas que utilizan la IA como su uso y la finalidad de los mismos.”²⁹²

Y a esta necesidad de “promover el respeto de los derechos humanos desde una perspectiva ética, solidaria y ecológica”²⁹³ mediante la elaboración de marcos y normas jurídicas garantistas, se suma la no menos importante tarea de establecer un marco de actuación ético que nos permita conocer la finalidad y los medios utilizados para conseguir el fin pretendido.

Desde este enfoque y apoyándonos en la clasificación que cita Pinazo, entendemos como “objetivo”, “un propósito fijado por la propia voluntad, que puede alcanzarse, establecerse y redefinirse secuencialmente - táctica o estratégicamente”-.

Mientras que el “fin”, refleja la idea instrumental, es un cometido impuesto desde fuera y que se debe cumplir continuamente y, por tanto, supone una prestación continuada”²⁹⁴..

Teniendo estas aclaraciones en cuenta, consideramos la conveniencia de que la ciudadanía reciba la adecuada información de manera veraz y transparente sobre el objeto, la finalidad y el propósito de la utilización de sus datos y por ello, también, “resulta imprescindible que la educación fomente el respeto a la diversidad, el diálogo intercultural y el sentido crítico frente a las informaciones y los algoritmos que condicionan nuestras decisiones por presentarnos una aparente verdad que se halla lejos de la realidad”²⁹⁵

Asistimos por tanto a una transformación digital con un impacto social sin precedentes

²⁹¹ Lasalle, J.M. Estamos adentrándonos en el riesgo de la creación en sus orígenes más profundos, entrevista en *Ethic*. Consultado en página web: <https://ethic.es/2024/05/entrevista-jose-maria-lasalle>

²⁹² De Bartolomé, J.C. Los nuevos valores y principios del S. XXI. El riesgo de la pérdida de la identidad humana, en *Revista Iberoamericana de Autogestión y Acción Comunal (RIDAA)*, n.º 81. Otoño 2023, Ed. INAUCO-UPV, pp.126-127

²⁹³ De Bartolomé, J.C. Los nuevos valores y principios del S. XXI. El riesgo de la pérdida de la identidad humana, en *Revista Iberoamericana de Autogestión y Acción Comunal (RIDAA)*, n.º 81. Otoño 2023, Ed. INAUCO-UPV, p. 137

²⁹⁴ Pinazo, J., *El impacto de las elecciones en la estructura de la Administración Pública, en Elecciones 2023. Un análisis multidimensional de las elecciones generales, autonómicas y municipales del año 2023*, Ed. Tirant lo Blanch, Valencia, 2024, p.43

²⁹⁵ De Bartolomé, J.C. Los nuevos valores y principios del S. XXI. El riesgo de la pérdida de la identidad humana, en *Revista Iberoamericana de Autogestión y Acción Comunal (RIDAA)*, n.º 81. Otoño 2023, Ed. INAUCO-UPV, p. 134

que altera las reglas del juego de espacios tradicionalmente regulados y ante unas dinámicas relacionales aún desconocidas.

A lo anteriormente señalado habría que añadir la necesidad de establecer acuerdos base enmarcados en unas normas éticas de obligado cumplimiento que contribuyan a asegurar un uso responsable de la IA y de crear mecanismos de control y supervisión que garanticen la buena praxis en la utilización de las nuevas tecnologías o de lo contrario, cabe la posibilidad de que podemos encontrarnos ante comportamientos con derivas ciertamente perjudiciales para la protección de los derechos humanos y de las libertades fundamentales.

En este proceso de cambio es fundamental, a nuestro entender, marcar las líneas rojas que no deben ser traspasadas, en aras a permitir una convivencia social, segura, efectiva y respetuosa con los derechos y libertades de las personas en las sociedades democráticas ya que como plantea Lasalle "si el desarrollo competitivo de la IA lo están haciendo dos súper potencias que están luchando por la hegemonía planetaria buscando tener empresas más competitivas, ejércitos más letales y gobiernos con mayor capacidad de control social, ¿cómo vamos a garantizar, a nivel estructural, el desarrollo de una IA para el bien común?"²⁹⁶

En estas coordenadas, consideramos relevante señalar algunas de las decisiones que marcaron la cumbre internacional impulsada por el Gobierno Federal de Alemania, la "Hamburg Sustainability Conference (HSC) 2025"²⁹⁷, que reunió a líderes de empresas, gobiernos, sociedad civil y organismos multilaterales con el compromiso común de establecer un marco de acción global para impulsar compromisos transformadores en torno a los Objetivos de Desarrollo Sostenible (ODS)²⁹⁸, figurando entre ellos la necesidad de marcos regulatorios sobre el desarrollo y uso ético de la IA.

En la mencionada Cumbre se firmó la "Hamburg Declaration on Responsible AI for the SDGs"²⁹⁹, con el compromiso de fomentar el desarrollo y el uso ético de la IA poniendo el foco sobre tres ejes: los derechos humanos, la equidad y la sostenibilidad.

²⁹⁶ Lasalle, J.M. *Estamos adentrándonos en el riesgo de la creación en sus orígenes más profundos*, entrevista en Ethic. Consultado en página web: <https://ethic.es/2024/05/entrevista-jose-maria-lassalle>

²⁹⁷ Hamburg Sustainability Conference (HSC) 2025, consultado en la página web de la ONU, <https://www.undp.org/events/undp-2025-hamburg-sustainability-conference>.

²⁹⁸ Ojetivos de Desarrollo Sostenible (ODS) consultado en la página web de la ONU, <https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/>

²⁹⁹ Hamburg Declaration on Responsible AI for the SDGs, consultado en la página web de la ONU: <https://www.undp.org/news/hamburg-declaration-responsible-ai-global-leaders-commit-responsible-ai-sustainable-development>

En esta Cumbre, se abordó la necesidad de establecer normativas estables que garanticen el buen uso de la IA, la seguridad de los consumidores y usuarios y el respeto de los derechos fundamentales.

Sin duda la aplicación de la IA puede contribuir a desarrollar avances de gran impacto en distintos ámbitos de la sociedad, pero supone al mismo tiempo la necesidad de articular mecanismos adecuados que permitan establecer como objetivo prioritario proteger los derechos humanos y preparar a todas las personas con los valores adecuados y las aptitudes necesarias para una colaboración eficaz entre las personas y las máquinas. Relacionar la IA con la educación en las políticas públicas educativas, contando con los diversos agentes sociales para la gobernanza, con el compromiso común del respeto a los textos internacionales y a los derechos humanos”³⁰⁰

En este escenario, la UNESCO elaboró la primera norma mundial sobre la ética de la IA: la "Recomendación sobre la ética de la inteligencia artificial", aplicable a los 193 Estados miembros que la componen³⁰¹.

En dicha recomendación se establecen diez principios básicos:

1. Proporcionalidad e inocuidad en el uso de sistemas de IA para alcanzar un objetivo legítimo
2. Seguridad y protección frente a riesgos y daños no deseados.
3. Derecho a la intimidad y protección de datos
4. Gobernanza y colaboración adaptativas y de múltiples partes interesadas desde el respeto al derecho internacional y la soberanía nacional.
5. Responsabilidad y rendición de cuentas. Los sistemas de IA deben ser auditables y trazables.
6. Transparencia y explicabilidad.
7. Supervisión y decisión humanas en lo que respecta a la asunción de responsabilidades éticas y jurídicas.
8. Sostenibilidad

³⁰⁰De Bartolomé, J.C. *Los nuevos valores y principios del S. XXI. El riesgo de la pérdida de la identidad humana*, en Revista Iberoamericana de Autogestión y Acción Comunal (RIDAA), n.º 81. Otoño 2023, Ed. INAUCO-UPV, p. 139

³⁰¹UNESCO, *Ética de la inteligencia artificial*, consultado en la página de la UNESCO: <https://www.unesco.org/es/artificial-intelligence/recommendation-ethics?hub=32618>

9. Sensibilización y educación de manera abierta y accesible que permita la participación de la ciudadanía.

10. Equidad y no discriminación

Se trata básicamente, como aclara De Bartolomé, “de un marco normativo, no vinculante y de carácter programático, para construir la superestructura jurídica. Recoge una serie de valores y principios para desarrollar prácticas de ésta tecnología saludables y no invasivas”³⁰²

En este sentido hay que reivindicar “la Declaración Universal de Derechos Humanos de 1948, la Carta Fundacional de 1945 de Naciones Unidas, así como los principios y valores fundamentales, contemplados en el artículo 2, del Tratado de la Unión Europea, y en nuestro caso español la vigencia del régimen constitucional de 1978, para garantizar la vida humana, la dignidad y la libertad del ser humano”³⁰³

2 La IA y sus relaciones con la ciudadanía

Los nuevos desafíos y oportunidades ante los que nos encontramos a medio y largo plazo, nos muestran a grandes rasgos la necesidad de establecer unas líneas de actuación estratégicas que de manera transversal garanticen el respeto a los derechos humanos y a las libertades fundamentales, ya que como afirma Lasalle, “por primera vez el ser humano se relacionará con una alteridad radical, con un otro con el que no puede empatizar y sin condicionantes morales. Ahí se va a producir un abismo real”³⁰⁴ y desde este prisma, consideramos la importancia de compatibilizar los diversos usos de la IA con una regulación que, garantice, refuerce y preserve el respeto efectivo de los derechos fundamentales de la ciudadanía.

³⁰²De Bartolomé, J.C. Los nuevos valores y principios del S-XXI. El riesgo de la pérdida de la identidad humana. RIDAA. Revista Iberoamericana de Autogestión y Acción Comunal, n.º 81. Ed. INAUCO – UPV. Otoño 2023, p. 137

³⁰³Canales, J.M. Notas sobre el transhumanismo versus la gobernanza democrática, en RIDAA. Revista Iberoamericana de Autogestión y Acción Comunal, n.º 82. Ed. INAUCO – UPV. Otoño 2024, p. 69

³⁰⁴Lasalle, J.M, Sin conciencia, la inteligencia artificial sigue el modelo de un psicópata. Entrevista en Innovaspain, 2025. Consultado en página web: <https://www.innovaspain.com/entrevista-jose-maria-lasalle/>

De la protección de los Derechos fundamentales

Muchos de los estudios e investigaciones que sobre el desarrollo y la aplicación de la IA han sido publicados, se centran en ensalzar las enormes oportunidades que los avances tecnológicos pueden aportar a las sociedades en distintos ámbitos, pero también muestran su inquietud ante los posibles riesgos y amenazas que puede suponer para los derechos fundamentales en las sociedades democráticas.

Hay que tener en cuenta como indica Cotino que “los algoritmos generalmente dependen de un consumo masivo de datos [...] y aunque se presenten como herramientas neutrales y objetivas, no están exentos de sesgos y errores. Estos sesgos fácilmente se proyectan en individuos o en colectivos generando tratamientos diferentes que pueden no estar suficientemente justificados.

Estos tratamientos diferenciados pueden tener consecuencias que son especialmente graves cuando se utilizan en ámbitos como el empleo, la educación, la concesión de créditos bancarios, las prestaciones, los contratos y ayudas sociales o la determinación de penas. Las discriminaciones algorítmicas, ya sean conscientes o no, pueden perpetuar y amplificar desigualdades estructurales preexistentes”³⁰⁵.

Por citar algún ejemplo, podríamos mencionar la aplicación de la IA en ámbitos relacionados con la salud. Por un lado, podríamos considerar el hecho de que contribuiría al avance y mejora de diagnósticos y tratamientos médicos enfocados al bienestar de los pacientes, pero, por otro lado, podría darse el caso de accesos indiscriminados a expedientes médicos, que, sin autorización o consentimiento de la persona interesada, pudieran ser utilizados con fines discriminatorios o excluyentes.

En este sentido, la OMS publicó en el año 2021 el primer informe mundial sobre IA aplicada a la salud y seis principios básicos rectores relativos a su concepción y utilización en el que se destaca que “guiándose por las leyes existentes y las obligaciones en materia de derechos humanos, así como por las nuevas leyes y políticas que incorporan principios éticos, los gobiernos, los proveedores y los diseñadores deben trabajar conjuntamente para abordar

³⁰⁵Cotino, L., *Encuesta sobre inteligencia artificial y derechos fundamentales. Revista Teoría y realidad constitucional*, n.º 55, 2025, consultado en la página web de la UNED: <https://uned.es/index.php/TRC/issue/view/1977>, p.17

las preocupaciones éticas y de derechos humanos en cada etapa de la concepción, desarrollo y despliegue de una tecnología basada en la IA”³⁰⁶.

Otro ejemplo que podríamos citar es el de la posible utilización de la herramienta del reconocimiento facial basado en la IA, ya que como afirma Ridaura "tiene innumerables ventajas, particularmente en el ámbito de la seguridad ya que puede coadyuvar al esclarecimiento de la investigación criminal y también se ha demostrado su utilidad en el reconocimiento de acosadores, en la lucha antiterrorista, o en la identificación de víctimas de catástrofes, entre otros [...] pero la ausencia de una ordenación jurídica con garantías puede conducir a escenarios de vigilancia masiva, afectando a derechos fundamentales que van más allá de la privacidad y la protección de datos, pues también puede perturbar la libertad de movimiento, la libertad ideológica, la de reunión, la presunción de inocencia, el derecho a la tutela judicial efectiva, o la prohibición de discriminación, por poner algunos ejemplos, ya que hay que tener en cuenta que en un "Estado democrático las intromisiones en la privacidad para preservar la seguridad demandan justificación y deben respetar criterios de idoneidad, necesidad y proporcionalidad”³⁰⁷

La desinformación

Las sociedades democráticas se encuentran ante otro desafío que es el de la desinformación y en este sentido, Suárez de Lezo asevera que “en el ecosistema digital contemporáneo, donde cada clic tiene valor y cada segundo de atención es mercancía, se extiende una amenaza soterrada: la consolidación de la industria de los falsos medios. Este fenómeno, impulsado por la lógica y opacidad de la publicidad programática y por la sofisticación de las herramientas de inteligencia artificial, no solo erosiona la confianza en la información, sino que compromete la viabilidad de los medios tradicionales”³⁰⁸

Para poder conocer con exactitud el tipo de reto que debe ser abordado y dadas sus amplias connotaciones, nos apoyamos en Cernada para distinguir, con carácter general” dos conceptos de

³⁰⁶Organización Mundial de la Salud, (OMS), *La OMS publica el primer informe mundial sobre inteligencia artificial (IA) aplicada a la salud y seis principios rectores relativos a su concepción y utilización*, consultado en página web: <https://www.who.int/es/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use>

³⁰⁷Ridaura, M. J. *El reconocimiento facial como herramienta securitaria*, consultado en *Retos de la inteligencia artificial para el Estado Social y Democrático de Derecho, Proyecto Retina-Der*, en la página web: <https://retina-der.uniovi.es/index.php/2024/12/17/el-reconocimiento-facia-como-herramienta-securitaria>, 2025.

³⁰⁸Suárez De Lezo, J, *El patrocinio de la desinformación y los falsos medios*, Informe consultado en la página web de Agenda Pública. Análisis de políticas públicas: <http://agendapublica.es/noticia/20010/desinformacion-patrocinio-fake-news>, 2025

desinformación.

En primer lugar, un concepto más estricto de desinformación que se caracterizaría por referirse a una información deliberadamente falsa que genera un perjuicio público [...] o bien, puede tratarse de un error por ausencia de voluntad de engaño u otras realidades, que se alinean con el ejercicio de la libertad de expresión o la propaganda”³⁰⁹

En estas coordenadas surge la necesidad de distinguir entre ambos conceptos y adaptar herramientas y mecanismos que garanticen una convivencia social respetuosa con los valores democráticos, eficaz en la protección y defensa de los mismos y mitigando los posibles perjuicios que pudieran ocasionarse por la utilización de las nuevas tecnologías y por el desarrollo de la IA.

En una sociedad globalizada como la actual, “la desinformación no conoce fronteras. Las mismas narrativas, apenas ajustadas, circulan en distintos países y contextos culturales. La cuestión migratoria es el ejemplo más claro. Las mismas fórmulas —miedo, invasión, reemplazo— se replican en Italia, España, Francia o Alemania con variaciones mínimas”³¹⁰

Es por tanto, indispensable, encontrar un modelo normativo que regule y garantice las nuevas relaciones generadas por las innovaciones tecnológicas y más concretamente por la IA, ya que como afirma Cernada, “en este nuevo ecosistema de información, cabe cuestionarse a quién corresponde el control de la desinformación.

Tres son los sujetos que pueden ejercer el control frente a este fenómeno: las plataformas o redes sociales, las Administraciones Públicas y los usuarios y la sociedad civil”³¹¹

De ahí la importancia de diseñar y desarrollar medidas tanto preventivas como correctivas que nos permitan conocer el cómo y el por qué del comportamiento de los algoritmos.

A este respecto, señala De Bartolomé, el Tribunal Constitucional de España “ha fijado una serie de criterios que deben tenerse en cuenta, como la veracidad, licitud y relevancia

³⁰⁹Cernada, R. *La transparencia de las plataformas digitales como garantía democrática en el control de la desinformación: una vuelta al Derecho público en clave algorítmica*, Revista Valenciana d’estudis autonòmics, 2023, n.º 68, p.156

³¹⁰Suárez De Lezo, J. *El patrocinio de la desinformación y los falsos medios*, Informe consultado en la página web de Agenda Pública. Análisis de políticas públicas: <http://agendapublica.es/noticia/20010/desinformacion-patrocinio-fake-news>, 2025

³¹¹Cernada, R. *La transparencia de las plataformas digitales como garantía democrática en el control de la desinformación: una vuelta al Derecho público en clave algorítmica*, Revista Valenciana d’estudis autonòmics, 2023, n.º 68, p.159

pública de la información, el consentimiento expreso del afectado (o, al menos, no presunto automático); el tiempo transcurrido desde la publicación: o la sensibilidad de los datos personales”³¹²

En el ámbito internacional, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) ha desarrollado una serie de principios mínimos que deberían cumplir los sistemas de IA³¹³.

Dichas recomendaciones identifican “cinco principios complementarios basados en valores para la administración responsable de la IA:

1. La IA debe **beneficiar a las personas y al planeta** impulsando el crecimiento inclusivo, el desarrollo sostenible y el bienestar.
2. Los sistemas de IA deben diseñarse **respetando el estado de derecho, los derechos humanos, los valores democráticos y la diversidad**, y deben incluir los mecanismos apropiados - como, por ejemplo, permitir la intervención humana cuando sea necesario- para garantizar una sociedad justa y equitativa.
3. Debe existir una **transparencia y divulgación responsable** en torno a los sistemas de IA, para garantizar que las personas entiendan sus resultados y puedan desafiarlos.
4. Los sistemas de IA deben **funcionar de manera sólida y segura** a lo largo de su ciclo de vida y los **riesgos potenciales deben evaluarse y gestionarse continuamente**.
5. Las **organizaciones e individuos que desarrollan, despliegan u operan sistemas de IA** deben ser **responsables de su correcto funcionamiento**, en base a los principios anteriormente descritos.

Tomando como referencia estos principios, la OCDE también proporciona **cinco recomendaciones a los gobiernos**:

1. **Facilitar la inversión pública y privada** en investigación y desarrollo, con el objetivo de estimular la innovación en IA de manera segura y fiable.
2. **Fomentar los ecosistemas de IA accesibles** con infraestructura y tecnologías digitales, y mecanismos para compartir datos y conocimientos.
3. **Asegurar un marco de políticas** que abra el camino para el despliegue de sistemas de IA fiable.
4. **Capacitar a las personas** con habilidades necesarias para la IA y apoyar a los trabajadores para una transición justa.

³¹²De Bartolomé, J.C Los nuevos valores y principios del S-XXI. El riesgo de la pérdida de la identidad humana. RIDAA. Revista Iberoamericana de Autogestión y Acción Comunal, n.º 81. Ed. INAUCO – UPV. Otoño 2023, p. 133

³¹³OCDE, Los principios de Inteligencia Artificial de la OCDE. Consultado en página web del Ministerio para la transformación digital y de la Función Pública del Gobierno de España <https://datos.gob.es/es/blog/los-principios-de-inteligencia-artificial-de-la-ocde>

5. **Cooperar a través de las fronteras y los sectores** para avanzar en la administración responsable de IA fiable”.

Así, en este nuevo escenario de revolución tecnológica se hace necesario establecer como objetivo prioritario la protección de los derechos humanos y en este sentido, asumir las nuevas realidades desde modelos tradicionales suponen un reto y un desafío, ya que se han de tener en cuenta numerosas variables, entre ellas, como hemos expuesto anteriormente, la seguridad y la protección de las personas.

De este modo, como afirma Oliver, “al igual que se regulan los alimentos, los medicamentos o los coches, debe regularse el uso de la inteligencia artificial. No la disciplina en sí, sino las aplicaciones concretas y sus impactos”.

Se trata, como indica Guterres de un “desafío social de primera magnitud, con implicaciones directas en derechos humanos, igualdad de oportunidades y justicia social³¹⁴”

3. La regulación normativa de la Inteligencia Artificial

La legislación europea

Con estos mimbres, la Comisión Europea tras presentar el Reglamento de Servicios Digitales (RSD)³¹⁵ y el Reglamento de Mercados Digitales³¹⁶, ha propuesto una ambiciosa legislación a nivel de la UE para defender los valores democráticos europeos.

Se trata de “una regulación enfocada sobre todo al control de los riesgos que plantea la IA, para la seguridad de consumidores y usuarios y del respeto a los derechos fundamentales, dirigida principalmente a proveedores y operadores, en el marco de la UE, pero con aspiraciones globales³¹⁷”.

³¹⁴Guterres, A, Secretario General de la ONU, *La gobernanza de la IA, un reto social y de derechos humanos*, en el *Diálogo Global sobre Gobernanza de la IA*. Información consultada en la página web del *Diario Responsable*:

<https://diarioresponsable.com/noticias/38592-la-gobernanza-de-la-ia-un-reto-social-y-de-derechos-humanos>

³¹⁵Reglamento de Servicios Digitales “regula los intermediarios y plataformas en línea, tales como mercados, redes sociales, plataformas de intercambio de contenidos, tiendas de aplicaciones y plataformas de viajes y alojamiento en línea. Su principal objetivo es prevenir las actividades ilegales y nocivas en línea y la difusión de desinformación. Garantiza la seguridad de los usuarios, protege los derechos fundamentales y crea un entorno de plataformas en línea justo y abierto”. Información consultada en la página web de la Comisión Europea: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_es

³¹⁶El Reglamento de Mercados Digitales es uno de los elementos centrales de la Estrategia Digital Europea y establece un conjunto de criterios objetivos claramente definidos para considerar a una gran plataforma en línea como «guardián de acceso» y garantiza que se comporte de manera justa en línea y deje margen para la disputabilidad.

Información consultada en la página web de la Comisión Europea: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_es

³¹⁷Ortega, A, *Hacia un régimen europeo de control de la Inteligencia Artificial*, consultado en la página web de la Revista del Real Instituto Elcano: <https://www.realinstitutoelcano.org/analisis/hacia-un-regimen-europeo-de-control-de-la-inteligencia-artificial/>

En el ámbito de la Unión Europea se ha “optado por un modelo de corregulación que parte del Plan de acción para la democracia europea y que se articula sobre dos normas fundamentales: el Reglamento de servicios digitales (DSA)³¹⁸ y la Ley de IA.³¹⁹

El nuevo modelo pivota sobre tres pilares básicos que se identifican con los sujetos con capacidad de ejercer funciones de control en la red: los usuarios, la colaboración civil y la cooperación internacional³²⁰

En lo que respecta a la DSA, Cotino afirma que “regula garantías y reclamaciones frente a las decisiones de las plataformas, y asigna un papel regulador significativo a las autoridades nacionales, denominadas “coordinadores de servicios digitales” que han de supervisar y organizar la implementación de estas normas. En España, la autoridad designada ha sido la CNMC³²¹”

Y con el fin de abordar los riesgos para los derechos fundamentales y la seguridad, la Comisión europea establece unas normas, a través de un Reglamento, por su carácter vinculante para todos los Estados miembros y en todos sus elementos que contiene los siguientes requisitos³²²:

1. Sistematizar los riesgos que pudieran derivarse de la aplicación de la IA: como riesgo inaceptable, alto riesgo, riesgo limitado o mínimo o ningún riesgo.

³¹⁸ La Ley de Servicios Digitales (DSA) es un nuevo conjunto de normas a escala de la UE para los servicios digitales que actúan como intermediarios para los consumidores y los bienes, servicios y contenidos. En el contexto de la DSA, los servicios digitales se refieren a los servicios intermediarios, como los proveedores de alojamiento, los mercados en línea y las redes sociales.

La DSA tiene como objetivo construir un mundo en línea más seguro y justo. Introducirá normas que protejan igualmente a todos los usuarios de la UE, tanto en lo que se refiere a bienes, contenidos o servicios ilícitos, como a sus derechos fundamentales.

Consultado en página web de la Comisión Europea: <https://digital-strategy.ec.europa.eu/es/faqs/digital-services-act-questions-and-answers>

³¹⁹ Ley de la UE, consultado en la página web del Parlamento Europeo:

<https://www.europarl.europa.eu/topics/es/article/20230601STO93804/ley-de-ia-de-la-ue-primer-norma-sobre-inteligencia-artificial>

³²⁰ Cernada, R. La transparencia de las plataformas digitales como garantía democrática en el control de la desinformación: una vuelta al Derecho público en clave algorítmica, *Revista Valenciana d'estudis autonòmics*, 2023, n.º 68, p.159

³²¹ Cotino, L. Entre bytes y democracia. Cómo la DSA de la UE armoniza la libertad de expresión y la lucha contra la desinformación. Consultado en la página web del Muy Ilustre Colegio de Abogados de la Abogacía de Madrid, Otrosí, n.º 2/VII 2024, 9ª Época: https://www.otrosi.net/wp-content/uploads/2024/07/Otrosi-no2_VII_2024.pdf, pp.42-44

³²² De Bartolomé, J.C Los nuevos valores y principios del S-XXI. El riesgo de la pérdida de la identidad humana. *RIDAA. Revista Iberoamericana de Autogestión y Acción Comunal*, n.º 81. Ed. INAUCO – UPV. Otoño 2023, pp. 145-147

2. Enumeración de una lista de aplicaciones de IA de alto riesgo.
3. Relacionar requisitos indubitados para los sistemas de IA para las aplicaciones de alto riesgo;
4. Realización de una evaluación exhaustiva de conformidad antes de que el sistema de IA se ponga en servicio o se comercialice.
5. Evaluación después de que dicho sistema de IA se comercialice
6. Establecer una gobernanza multinivel, europea y nacional.

Con ello, afirma De Bartolomé, “se pretende establecer un régimen armonizado, coherente y homogéneo para tratar las reclamaciones de responsabilidad de los consumidores por los daños causados por productos y servicios basados en IA”

Y en esta línea de actuación, Ortega señala que uno de los requisitos más básicos planteado por la Comisión, es que “la IA debe de ser transparente, lo que supone poder reconstruir cómo y por qué se comporta de una determinada manera. Quienes interactúen con esos sistemas deben de saber que se trata de IA así como qué personas son sus responsables”³²³

En resumen, la UE marca un hito histórico con su norma jurídica de regulación de la IA y en la que a través de sus 12 títulos principales, busca “garantizar que la IA desarrollada y utilizada en la UE sea segura, fiable y centrada en la protección de los derechos fundamentales de las personas”³²⁴.

Desde este enfoque, la Ley de IA clasifica los sistemas de IA en función de su riesgo, categorizando los sistemas de IA en cuatro niveles: riesgo mínimo, riesgo específico de transparencia, riesgo alto y riesgo inaceptable. Por ejemplo, los chatbots deberán informar claramente a los usuarios que están interactuando con una máquina, mientras que sistemas considerados de alto riesgo, como los utilizados en la contratación o en la evaluación de créditos, deberán cumplir con estrictos requisitos de calidad y supervisión. [...] Además,

³²³Ortega. A, *Hacia un régimen europeo de control de la Inteligencia Artificial*, consultado en la página web de la Revista del Real Instituto Elcano: <https://www.realinstitutoelcano.org/analisis/hacia-un-regimen-europeo-de-control-de-la-inteligencia-artificial/>

³²⁴Ley de Inteligencia Artificial de la UE, consultada en la página web: <https://artificialintelligenceact.eu/es/>

introduce normas específicas para los modelos de IA de uso general, garantizando la transparencia y abordando posibles riesgos sistémicos.³²⁵

³²⁵Ley europea para regular la IA. Información consultada en la página de la Agencia Digital de Andalucía. Desarrollo de Servicios Digitales: <https://desarrollo.juntadeandalucia.es/node/2593>

4. Conclusiones

El cambio que supone la aplicación de sistemas de IA en la vida cotidiana de las personas ha generado por un lado altas expectativas, pero por otro, desde algunos sectores se plantean los límites legales y éticos de su utilización.

Este proceso de cambio que se produce en las sociedades democráticas actuales sobre el diseño, el desarrollo, la aplicación, la gestión y el tratamiento de los datos utilizados a través de mecanismos que aplican la IA, requiere de una amplia regulación que enmarque, oriente y establezca los principios rectores de su utilización.

En este sentido la Unión Europea, es pionera en el mundo en la creación de un marco jurídico que garantiza, asegura y protege, a través de la Ley de IA, los derechos y libertades fundamentales de la ciudadanía ante los retos, desafíos y peligros que puede entrañar la aplicación de las nuevas tecnologías y particularmente la IA.

La protección de los derechos fundamentales, arduamente conquistados en las sociedades democráticas actuales, requiere además de instrumentos de control de la utilización de sistemas de IA en las distintas esferas de actuación de sociedades cada vez más polarizadas.

Desde esta perspectiva, consideramos relevante destacar que diversas organizaciones internacionales han puesto de manifiesto la importancia de sumar a los fundamentos jurídicos, normas éticas que regulen las relaciones sociales en los ecosistemas tecnológicos actuales, que permitan delimitar su impacto y que garanticen una gestión de datos responsable y transparente.

En este sentido, Oliver nos recuerda que “Está en nuestras manos decidir qué tipo de inteligencia artificial queremos. [...] Como sociedad, debemos decidir qué desarrollo queremos: uno que contribuya al bienestar, al progreso y que beneficie a todas las personas, no solo a unas pocas. Ni a costa del planeta³²⁶”.

³²⁶Oliver, N. *La inteligencia artificial nos fascina porque nos hace sentir como dioses*, entrevista en Santander Women Now, consultado en la página web: <https://www.mujerhoy.com/womennow/santander-womennow-nuria-oliver-experta-ia-desafios-oportunidades-inteligencia-artificial-20250619110841-nt.html>

Bibliografía

Canales, J.M. *Notas sobre el transhumanismo versus la gobernanza democrática*, en RIDAA. Revista Iberoamericana de Autogestión y Acción Comunal, n.º 82. Ed. INAUCO – UPV. Otoño 2024.

Cernada, R. *La transparencia de las plataformas digitales como garantía democrática en el control de la desinformación: una vuelta al Derecho público en clave algorítmica*, Revista Valenciana d'estudis autonòmics, 2023, n.º 68.

Cotino, L. *Implantando la transparencia y los registros públicos de algoritmos. La experiencia pionera en la Comunidad Valenciana*, Revista Valenciana d'Estudis Autonòmics, 2023, n.º 68.

De Bartolomé, J.C. *Los nuevos valores y principios del S. XXI. El riesgo de la pérdida de la identidad humana*, en Revista Iberoamericana de Autogestión y Acción Comunal (RIDAA), n.º 81, otoño 2023, Ed. INAUCO-UPV.

Pinazo, J., *El impacto de las elecciones en la estructura de la Administración Pública, en Elecciones 2023. Un análisis multidimensional de las elecciones generales, autonómicas y municipales del año 2023*, Ed. Tirant lo Blanch, Valencia, 2024.

Webgrafía y otras fuentes digitales

Cotino, L., *Encuesta sobre inteligencia artificial y derechos fundamentales*, en la revista *Teoría y realidad constitucional*, n.º 55, 2025, consultado en la página web de la UNED: <https://uned.es/index.php/TRC/issue/view/1977>

Cotino, L. *Entre bytes y democracia. Cómo la DSA de la UE armoniza la libertad de expresión y la lucha contra la desinformación*. Consultado en la página web del Muy Ilustre Colegio de Abogados de la Abogacía de Madrid, Otrosí, n.º 2/VII 2024, 9ª Época: https://www.otrosi.net/wp-content/uploads/2024/07/Otrosi-no2_VII_2024.pdf

Declaración Universal de los Derechos Humanos. Consultado en la página web de las Naciones Unidas: <https://www.un.org/es/about-us/universal-declaration-of-human-rights>

Guterres, A, Secretario General de la ONU, *La gobernanza de la IA, un reto social y de derechos humanos*, en el Diálogo Global sobre Gobernanza de la IA. Información consultada en la página web del Diario Responsable: <https://diarioresponsable.com/noticias/38592-la-gobernanza-de-la-ia-un-reto-social-y-de-derechos-humanos>

Hamburg Declaration on Responsible AI for the SDGs, consultado en la página web de la ONU: <https://www.undp.org/news/hamburg-declaration-responsible-ai-global-leaders-commit-responsible-ai-sustainable-development>

Hamburg Sustainability Conference (HSC) 2025, consultado en la página web de la ONU: <https://www.undp.org/events/undp-2025-hamburg-sustainability-conference>

Lasalle, J.M. *Estamos adentrándonos en el riesgo de la creación en sus orígenes más profundos*, entrevista en Ethic. Consultado en página web: <https://ethic.es/2024/05/entrevista-jose-maria-lasalle>

Lasalle, J.M, *Sin conciencia, la inteligencia artificial sigue el modelo de un psicópata*. Entrevista en Innovaspain, 2025, consultado en página web: <https://www.innovaspain.com/entrevista-jose-maria-lasalle/>

Ley de la UE, consultado en la página web del Parlamento Europeo: <https://www.europarl.europa.eu/topics/es/article/20230601STO93804/ley-de-ia-de-la-ue-primera-normativa-sobre-inteligencia-artificial>

Ley europea para regular la IA. Información consultada en la página de la Agencia Digital de Andalucía. Desarrollo de Servicios Digitales: <https://desarrollo.juntadeandalucia.es/node/2593>

La Ley de Servicios Digitales (DSA), consultado en página web de la Comisión Europea: <https://digital-strategy.ec.europa.eu/es/faqs/digital-services-act-questions-and-answers>

Objetivos de Desarrollo Sostenible (ODS) consultado en la página web de la ONU: <https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/>

OCDE, Los principios de Inteligencia Artificial de la OCDE, consultado en página web del Ministerio para la transformación digital y de la Función Pública del Gobierno de España <https://datos.gob.es/es/blog/los-principios-de-inteligencia-artificial-de-la-ocde>

Oliver N., *¿La pluma inteligente?*, consultado en la página web de la Real Academia de Ingeniería: <https://www.raing.es/la-pluma-inteligente-por.núria-oliver/> (artículo publicado en el diario ABC el 14 de abril de 2023)

Oliver, N. *La inteligencia artificial nos fascina porque nos hace sentir como dioses*, entrevista en Santander Women Now, consultado en la página web: <https://www.mujerhoy.com/womennow/santander-womennow-nuria-oliver-experta-ia-desafios-oportunidades-inteligencia-artificial-20250619110841-nt.html>

Organización Mundial de la Salud, (OMS), información consultada en página web: <https://www.who.int/es/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use>

Ortega, A. *Hacia un régimen europeo de control de la Inteligencia Artificial*, consultado en página web del Real Instituto Elcano: <https://www.realinstitutoelcano.org/analisis/hacia-un-regimen-europeo-de-control-de-la-inteligencia-artificial/>

Reglamento de Mercados Digitales, información consultada en la página web de la Comisión Europea: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_es

Reglamento de Servicios Digitales, información consultada en la página web de la Comisión Europea: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_es

Ridaura, M. J. El reconocimiento facial como herramienta securitaria, consultado en *Retos de la inteligencia artificial para el Estado Social y Democrático de Derecho, Proyecto Retina-Der*, en la página web: <https://retina-der.uniovi.es/index.php/2024/12/17/el-reconocimiento-facia-como-herramienta-securitaria>, 2025.

Suárez De Lezo, J, *El patrocinio de la desinformación y los falsos medios*, Informe consultado en la página web de *Agenda Pública. Análisis de políticas públicas*: <https://agendapublica.es/noticia/20010/desinformacion-patrocinio-fake-news>, 2025

UNESCO, *Ética de la inteligencia artificial*, consultado en la página de la UNESCO: <https://www.unesco.org/es/artificial-intelligence/recommendation-ethics?hub=32618>

Actas del VIII Congreso sobre Derechos Humanos

Derechos Humanos e Inteligencia Artificial

Celebrado los días 16 y 17 de octubre de 2025 en el Ilustre Colegio de Abogados de Valencia.

Dirección académica y comité científico: Vicente Bellver Capella, Ángeles Solanes Corella y Jorge Sebastián Lozano.

Comité organizador:

- **Fundación Mainel:** Cristina Celda, Maria José Medialdea, Inmaculada Zarranz.
- **Ilustre Colegio de Abogados de Valencia.**

Organizado por la Fundación Mainel



Con el patrocinio del Ayuntamiento de Valencia y del Ministerio de Asuntos Exteriores, Unión Europea y Cooperación



Y la colaboración de:

