informática y DERECHO

2ª Época

informática y DERECHO

2ª Época

Revista Iberoamericana de Derecho Informático

(Primer Semestre 2026- n.º 18)

fiadi

FEDERACIÓN IBEROAMERICANA
DE ASOCIACIONES DE DERECHO
E INFORMÁTICA

Las opiniones expresadas en los artículos de este volumen son de exclusiva responsabilidad de sus autores

Fundación
de Cultura
Universitaria

ISSN: 2530-4496

Editorial Fundación de Cultura Universitaria
25 de Mayo 553 - Tel. 2 916 11 52
CP 11.000 Montevideo - Uruguay
ediciones@fcu.edu.uy
www.fcu.edu.uy

ÍNDICE

INDEX

Tema central:
Inteligencia Artificial, Derecho y Sociedad

PRESENTACIÓN

Los artículos aceptados para publicación en la Revista FIADI n.º 18 fueron organizados en torno a seis grandes ejes temáticos que articulan el debate jurídico contemporáneo sobre la inteligencia artificial:

Eje 1 - Gobernanza y regulación

«De la regulación a la implementación de la IA en la Administración Pública y la construcción de una gobernanza algorítmica», Verónica Juliana Caicedo Buitrago.

Eje 2 - Protección de datos personales

«A busca pela adequação: Transferência internacional de dados UE-Brasil pós jurisprudência Schrems», Francisco Pedro Gonçalves Carvalho de Sousa.

«Contratos de uso de IA generativa y ley aplicable: Límites del Reglamento Roma I en el entorno digital», Carlos Domínguez Padilla.

Eje 3 - Derechos fundamentales

«Impacto de la IA en la protección jurídica de los derechos fundamentales», M.ª Olivia Galán Azofra.

«Portadores de derechos fundamentales en el entorno de las manifestaciones emergentes de la IA», Federico César Lefranc Weegan.

«Justicia juvenil algorítmica: desafíos de la evidencia actuarial y el derecho a la defensa», Pedro Luis Bracho-Fuenmayor.

Eje 4 - Responsabilidad y riesgo algorítmico

«Responsabilidad de la empresa y compliance digital: el defecto de organización en la era algorítmica», José Gregorio Pumarejo Luchón.

«La trascendental condena a Meta y YouTube en Los Ángeles por la configuración de sus patrones oscuros», Diego Fierro Rodríguez.

Eje 5 - Administración Pública y justicia

«Cuestiones en torno a la aptitud de la IA para acelerar el proceso penal en supuestos de multirreincidencia desde la Fiscalía de Barcelona», Laura Echarri Nicolás y Diego Fierro Rodríguez.

«Uso eficiente de la IA en la evaluación probatoria del proceso administrativo sancionador: el caso peruano», Carmen Milagros Velarde Koechlin.

«Inteligencia artificial en la impartición de justicia en América Latina», Olivia Andrea Mendoza Enríquez.

«IA para la vigilancia ambiental: un modelo de crowdsourcing e IA con incentivos para la protección de áreas naturales protegidas», Vázquez de la Paz, Segura Cardona, Barba González y Muñoz Ayala.

Eje 6 - Transparencia y explicabilidad e impacto de la IA generativa

«Fundamentação jurídica e sistemas de IA: da valoração à explicabilidade contínua», Bergamini, Taboada y Soares.

«Explicabilidad y trazabilidad en la IA generativa en el ecosistema jurídico: hacia un modelo de gobernanza orientado a derechos», Iván Miguel García-López y Laura Trujillo-Liñán.

Asimismo, en la Revista FIADI n.º 18 incorporamos, además las ponencias de investigación que fueron premiadas en la novena edición de los Premios Valentín Carrascosa en la categoría de estudiantes, que reconocen la excelencia académica emergente en formación dentro del ámbito del derecho y las nuevas tecnologías:

1.er Premio: «Inteligencia artificial generativa en la rama judicial: desafíos para la función jurisdiccional y garantías del debido proceso», de Shellen Sharay Baez Diaz (Colombia).

2.º Premio: «Los sistemas de inteligencia artificial, ¿sujeto u objeto en la relación jurídica civil? Impacto en los derechos humanos», de Vania Jacomino González (Cuba).

3.er Premio: «Inteligencia Artificial y Ética: Desafíos para los Derechos Humanos», de Bárbara Rosa Chinea Cárdenas (Cuba).

Finalmente, el desglose del Panorama de la Revista n.º 18 en cifras quedó así: son en total 17 artículos publicados, entre investigaciones originales y trabajos premiados; seis ejes temáticos que estructuran el debate jurídico sobre IA; tres premios otorgados en ponencias de categoría estudiante de la novena edición de los Premios Valentín Carrascosa y tenemos varios países de iberoamérica representados: España, Portugal, México, Perú, Brasil, Venezuela, Colombia, Cuba, entre otros.

Tres pilares editoriales definen la identidad de la Revista FIADI y garantizan su relevancia como referente jurídico en el ámbito de la inteligencia artificial, el derecho y la sociedad en el espacio iberoamericano:

1. Investigación rigurosa con artículos académicos de alto nivel.

2. Impacto en políticas públicas con influencia en regulación y toma de decisiones.
3. Perspectiva y proyección con análisis comparado entre países.

La Revista FIADI se consolida así como un espacio de reflexión jurídica rigurosa y plural sobre los desafíos que la inteligencia artificial plantea al derecho, la sociedad y los derechos fundamentales en el contexto iberoamericano y global.

Prof. Dr. José Heriberto García Peña

Director Académico

PRÓLOGO

Es un alto honor iberoamericano prologar el n.º 18 de la Revista FIADI, segunda época, que lleva adelante el Prof. José Heriberto García Peña, vicepresidente de Investigación e Innovación de la institución, como parte de sus funciones y con el apoyo de un equipo académico y miembros del equipo editorial en la FIADI como la profesora Lucana Estevez.

La revista es una obra intelectual de la FIADI que se publica semestralmente cada año con el apoyo, participación y financiamiento de sus miembros, habilita un espacio académico único de especialidad en Informática y Derecho para abogados de diferentes latitudes que escriben y comparten sus conocimientos e investigaciones, es un texto útil para actualizarnos de las novedades jurídico - informáticas en iberoamérica y dejar huella de esa etapa de la historia.

Es un momento histórico el que vivimos, y hace bien mirar en retrospectiva para comprender el significativo avance tecnológico y su impacto en la ciencia del Derecho actual. El profesional con más experiencia viene de una generación de la *sociedad de la información y del conocimiento* —tercera ola— donde el abogado deja de usar la máquina de escribir para usar la computadora, donde estudia no sólo en bibliotecas físicas sino también en bibliotecas digitales, y donde accede a bases de datos jurídicas de jurisprudencia, doctrina y legislación para realizar sus investigaciones, trabajo jurídico y organizar sus litigios; hablamos de una sociedad donde aprender a buscar en Google, saber manejar Microsoft Office (Word, Power Point y Excel) y el manejo de una que otra base de datos jurídica eran suficientes para decir que se estaba al día tecnológicamente.

En noviembre de 2022 ocurre un fenómeno no visto desde hace décadas: un cambio de sociedad, pasamos a una sociedad de la inteligencia artificial (IA) y del *machine learning* (ML), este hito es representado por el lanzamiento público de ChatGPT de la empresa OpenAI, esto es un cambio trascendental en la historia de la humanidad puesto que la inteligencia, esa capacidad de comprender problemas y darles solución, es sistematizada para generar textos artificiales, y progresivamente también de forma multimodal imágenes y videos; recordemos el viejo aforismo latino *Ubi societas, ibi ius* que se traduce como «donde hay sociedad, hay derecho», esto significa que el derecho es inherente a cualquier grupo humano organizado, que ante una sociedad nueva el derecho justifica su existencia poniendo la ley y el orden necesarios para la convivencia, en este sentido, esta nueva sociedad marca un escenario doctrinal jurídico de enorme envergadura, terreno fértil para conjeturar, crear, idear y en definitiva hacer producción intelectual en esta nueva materia como es el derecho de la inteligencia

artificial como una subrama del derecho informático. En este punto el abogado debe formarse en desarrollar habilidades de legal *prompting*, con mucha práctica y esfuerzo, para participar de un mercado más competitivo con potenciales capacidades más inteligentes de la práctica jurídica; desde el lanzamiento de la IA generativa todos los abogados se han desactualizado de forma instantánea y es un volver a empezar, hace bien entonces prestar atención al primer mandamiento del decálogo del abogado del profesor Eduardo Couture, que dice: «Estudia. El derecho se transforma constantemente. Si no sigues sus pasos, serás cada día un poco menos abogado».

Esta Revista n.º 18 de la FIADI, precisamente es rica en contenido en esta sociedad de la IA y del ML con sus diversos artículos sobre IA que abordan la regulación e implementación de la IA en la administración pública y la gobernanza algorítmica, la protección jurídica de los derechos fundamentales, el *compliance* digital en la era algorítmica, los contratos de uso de inteligencia artificial generativa y la ley aplicable, la explicabilidad y trazabilidad de los sistemas de IA, la IA en el proceso judicial y en la impartición de justicia, la evaluación probatoria en el proceso administrativo sancionador, la justicia juvenil algorítmica, la vigilancia ambiental mediante *crowdsourcing* e IA y los desafíos éticos y de derechos humanos vinculados a los sistemas de IA.

La historia continúa sin freno y para marzo de 2023, con el lanzamiento de AutoGPT, una plataforma de IA de código abierto que permite a los usuarios automatizar proyectos de varios pasos y flujos de trabajo complejos con agentes de IA basados en el modelo de lenguaje grande (LLM) GPT-4 de OpenAI, ingresamos a una nueva sociedad más avanzada que la de la IA generativa —basada en *prompting* y de naturaleza reactiva—, la llamaremos la *sociedad agéntica*, está tomando forma aceleradamente y su característica esencial es su capacidad de ser autónoma en su toma de decisiones lo cual la hace proactiva, estamos frente a un punto de inflexión, porque las herramientas de IA generativa, empezando por los chats, están desarrollando capacidades agénticas, por ejemplo, en una conversación con ChatGPT, esta podría usar dichas capacidades razonando sobre el objetivo, decidiendo qué herramienta usar según la tarea, por ejemplo búsqueda web, lectura de archivos, creación de documentos, calendario, correo o análisis de datos, ejecutando esas herramientas cuando corresponde, integrando los resultados y devolviendo una respuesta final. Esto coloca a la ciencia jurídica en un nuevo momento a tanta velocidad que muchos análisis y trabajos jurídicos aún se concentran sólo en la IA generativa y pierden de vista su rápida evolución hacía la IA agéntica dejando de lado análisis jurídicos en este nivel. La orquestación de agentes y su uso, por ejemplo, en el campo militar es desconcertante, cuando aplican los LLM a agentes e implementan armas autónomas letales esto pone en un *shock* jurídico a los más célebres juristas por la asignación de responsabilidades, atribución de culpa, la ausencia de intervención humana, entre otros temas altamente complejos y controversiales, que se deben repensar, aquí estudiar no es suficiente, es necesario reinventar figuras jurídicas, ajustar conceptos tradicionales y hasta crear nuevas definiciones del derecho, hacia nueva teoría general del derecho más cercano a lo sintético.

Finalmente, no quiero dejar un tema crítico de la academia e investigación científica en general pero en particular en la jurídica en estas sociedades que surgen, emergen y conviven unas con otras, y es el de los datos sintéticos que son precisamente aquellos datos artificiales generados para imitar los datos reales, como el código en el software, los párrafos de argumentos y explicaciones en un libro o artículo científico, cláusulas contractuales o inclusive nuevas interpretaciones doctrinales, todo ello puede ser artificial, y se producen mediante técnicas de inteligencia artificial como el *deep learning* e IA generativa.

En una etapa de datos sintéticos, hay desafíos muy serios a momento de publicar una revista científica, porque es muy difícil determinar qué es una idea original de un autor cuyo contenido es escrito por IA aunque revisado por su autor —nueva categoría—, de una idea y contenido generados por IA completamente, y lo que cada vez será menos frecuente ver ideas y contenidos escritos íntegramente por seres humanos. De hecho cabe la pregunta si el mundo jurídico académico debería exigir a los escritores al igual que a los litigantes en algunos foros y jurisdicciones en algunos países una certificación de autoría no sólo de escritos sino también de tesis y trabajos académicos de tal modo que puedan certificar que no usaron IA ni en la idea ni en el contenido, mientras tanto la reputación profesional y la buena fe del abogado es suficiente aún.

Con estos humildes aportes, quiero invitarlos a leer y disfrutar de las emocionantes palabras y párrafos de cada uno de los artículos de esta Revista, la FIADI está comprometida con el desarrollo científico y la academia en temas de derecho e informática desde hace más de 40 años y lo seguirá haciendo con las nuevas generaciones.

Ariel Agramont Loza

Managua, 19 de junio de 2026

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 20-45

DE LA REGULACIÓN A LA IMPLEMENTACIÓN DE LA INTELIGENCIA ARTIFICIAL EN LA ADMINISTRACIÓN PÚBLICA Y LA CONSTRUCCIÓN DE UNA GOBERNANZA ALGORÍTMICA

FROM REGULATION TO THE IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE IN PUBLIC ADMINISTRATION AND THE CONSTRUCTION OF ALGORITHMIC GOVERNANCE

Veronica Juliana Caicedo Buitrago

Facultad de Business & Tech, Universidad Alfonso X el Sabio

Resumen

La incorporación de sistemas de inteligencia artificial en la Administración Pública constituye un elemento central de la transformación digital contemporánea, al incidir en la adopción de decisiones automatizadas con impacto en derechos fundamentales y en el ejercicio del poder público. En este contexto, la inteligencia artificial plantea desafíos jurídicos que trascienden lo tecnológico, exigiendo modelos de gobernanza que garanticen transparencia, control y rendición de cuentas.

El presente trabajo analiza el tránsito desde la regulación hacia la implementación efectiva de la inteligencia artificial en el sector público, tomando como referencia el marco normativo de la Unión Europea, en particular el Reglamento de Inteligencia Artificial y su interacción con el Reglamento General de Protección de Datos. Se examinan los principales problemas derivados de la opacidad algorítmica, la falta de auditorías efectivas y la compleja atribución de responsabilidad en decisiones automatizadas.

Asimismo, se incorpora una perspectiva comparada con América Latina, donde la adopción de estas tecnologías avanza sin marcos consolidados de gobernanza. Finalmente, se propone un modelo basado en *accountability, compliance* digital y supervisión humana significativa.

Palabras clave

Inteligencia artificial, gobernanza algorítmica, responsabilidad jurídica, Administración Pública, auditoría algorítmica

Abstract

The incorporation of artificial intelligence systems in Public Administration represents a central element of contemporary digital transformation, particularly due to its impact on automated decision-making processes affecting fundamental rights and the exercise of public power. In this context, artificial intelligence raises legal challenges that go beyond technical considerations, requiring governance models that ensure transparency, control and accountability.

This paper analyses the transition from regulation to the effective implementation of artificial intelligence in the public sector, taking as a reference the European Union regulatory framework, particularly the Artificial Intelligence Regulation and its interaction with the General Data Protection Regulation. It examines key issues related to algorithmic opacity, the lack of effective auditing mechanisms and the complex attribution of responsibility in automated decision-making.

In addition, a comparative perspective with Latin America is incorporated, where the adoption of these technologies is advancing without consolidated governance frameworks. Finally, the paper proposes a model based on accountability, digital compliance and meaningful human oversight.

Keywords

Artificial intelligence, algorithmic governance, legal responsibility, Public Administration, algorithmic auditing

I. Introducción

La progresiva incorporación de sistemas de inteligencia artificial en la Administración Pública se inscribe en el marco más amplio de la transformación digital del sector público, caracterizada por la automatización de procesos decisionales y la utilización intensiva de datos en la gestión de servicios públicos. Este fenómeno, lejos de constituir una mera innovación tecnológica, incide directamente en el ejercicio del poder público, en la medida en que condiciona la adopción de decisiones con impacto en derechos fundamentales y en la relación entre la Administración y la ciudadanía.

En este contexto, la inteligencia artificial plantea desafíos jurídicos que trascienden el ámbito técnico, al introducir nuevas formas de opacidad, complejidad y autonomía en los procesos decisionales. La creciente utilización de sistemas algorítmicos en la toma de decisiones administrativas evidencia la existencia de una tensión estructural entre, por un lado, los principios clásicos del Derecho administrativo, basados en la transparencia, la motivación y el control, y, por otro, las características propias de los sistemas de inteligencia artificial, especialmente aquellos basados en técnicas de aprendizaje automático.

A ello se suma la existencia de una brecha significativa entre el desarrollo de marcos normativos avanzados y su efectiva implementación en contextos reales. Si bien el modelo europeo ha avanzado de forma notable en la construcción de un sistema regulatorio orientado a la gestión de riesgos, la operatividad de estos instrumentos se enfrenta a limitaciones derivadas de la complejidad técnica de los sistemas, de la falta de capacidades institucionales y de la dificultad de articular mecanismos de control efectivos.

La relevancia del problema radica, por tanto, en la necesidad de garantizar que la incorporación de sistemas de inteligencia artificial en la Administración Pública se realice de manera compatible con los principios del Estado de Derecho y con la protección de los derechos fundamentales. En este sentido, la cuestión no se limita a la existencia de normas jurídicas, sino que se proyecta sobre su capacidad real para ser implementadas en entornos caracterizados por la automatización y la toma de decisiones a gran escala.

El presente trabajo tiene como objetivo analizar el tránsito desde la regulación hacia la implementación efectiva de la inteligencia artificial en la Administración Pública, tomando como referencia el marco normativo de la Unión Europea y su interacción con otros instrumentos relevantes. Asimismo, se pretende identificar los principales desafíos asociados a la opacidad algorítmica, la atribución de responsabilidad y la insuficiencia de los mecanismos tradicionales de control, para, finalmente, proponer un modelo de gobernanza algorítmica basado en la integración de elementos jurídicos, técnicos y organizativos.

Desde el punto de vista metodológico, la investigación se articula a partir de un enfoque jurídico-normativo, basado en el análisis sistemático de las principales fuentes del Derecho europeo en materia de inteligencia artificial, protección de datos y ciberseguridad. A ello se añade un enfoque doctrinal, sustentado en la revisión de la literatura académica especializada, así como una aproximación comparada que incorpora la realidad latinoamericana, con el fin de identificar diferencias en los modelos de gobernanza y en su grado de institucionalización.

El trabajo se estructura en cuatro apartados. En primer lugar, se examina el marco normativo aplicable a la inteligencia artificial en el contexto europeo. En segundo lugar, se analizan los principales desafíos derivados de su implementación en la Administración Pública. En tercer lugar, se estudian las implicaciones en materia de responsabilidad y control de los sistemas algorítmicos. Finalmente, se propone un modelo de gobernanza algorítmica orientado a garantizar una aplicación efectiva y respetuosa con los principios del ordenamiento jurídico.

II. Marco normativo de la inteligencia artificial

La irrupción de la inteligencia artificial en el ámbito de la Administración Pública ha motivado la configuración progresiva de un entramado normativo complejo, caracterizado por su naturaleza multinivel, su heterogeneidad y su carácter evolutivo. Lejos de constituir un sistema cerrado, el marco regulatorio de la inteligencia artificial en el contexto europeo se articula a partir de una combinación de normas vinculantes, instrumentos de *soft law* y estándares técnicos que, en conjunto, pretenden dar respuesta a los desafíos derivados de la automatización de decisiones con impacto jurídico.

En este sentido, el punto de partida viene determinado por la aprobación 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), que constituye el primer intento sistemático de regulación integral de la inteligencia artificial desde una perspectiva basada en el riesgo. Este instrumento introduce una clasificación de los sistemas de inteligencia artificial en función de su nivel de riesgo y los clasifica como inaceptable, alto, limitado y mínimo, estableciendo obligaciones diferenciadas para los operadores en cada categoría, con especial incidencia en aquellos sistemas utilizados en el ámbito de la Administración Pública y la justicia[1].

El enfoque basado en el riesgo adoptado por el RIA responde a la necesidad de equilibrar la promoción de la innovación tecnológica con la protección de los derechos fundamentales, configurando un modelo regulatorio que no prohíbe la inteligencia artificial en sí misma, sino determinados usos considerados incompatibles con los valores de la Unión Europea (en adelante UE). No obstante,

1 Unión Europea, *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*.

este planteamiento ha sido objeto de crítica doctrinal[2], en la medida en que su eficacia depende en gran medida de la correcta identificación de los sistemas de alto riesgo y de la capacidad real de los Estados miembros para supervisar su implementación.

Ahora bien, el RIA no opera de forma aislada, sino que se integra en un ecosistema normativo más amplio en el que destacan, de manera particular, el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE[3] (en adelante RGPD), y la normativa sectorial en materia de ciberseguridad, como la Directiva (UE) 2022/2555 del Parlamento Europeo y del Consejo, de 14 de diciembre de 2022, relativa a las medidas destinadas a garantizar un elevado nivel común de ciberseguridad en toda la Unión, por la que se modifican el Reglamento (UE) n.º 910/2014 y la Directiva (UE) 2018/1972 y por la que se deroga la Directiva (UE) 2016/1148[4] (en adelante Directiva NIS2). En efecto, la inteligencia artificial se nutre esencialmente de datos, lo que implica que gran parte de los riesgos asociados a su utilización se encuentran estrechamente vinculados con el tratamiento de información personal y, en consecuencia, con la aplicación del régimen de protección de datos.

En este contexto, el RGPD adquiere una relevancia central, especialmente en lo relativo a las decisiones automatizadas y al *profiling*, regulados en su artículo 22, así como en las exigencias de transparencia, minimización de datos y responsabilidad proactiva del responsable del tratamiento. La interacción entre el RGPD y el RIA plantea, sin embargo, importantes desafíos interpretativos, en la medida en que ambos instrumentos comparten ámbitos materiales parcialmente coincidentes, pero responden a lógicas regulatorias distintas[5].

Por su parte, la Directiva NIS2 introduce obligaciones reforzadas en materia de gestión de riesgos y seguridad de redes y sistemas de información, lo que resulta particularmente relevante en el contexto de sistemas de inteligencia artificial utilizados por entidades públicas. La seguridad de los sistemas algorítmicos se configura como un elemento indispensable para garantizar no solo la continuidad de los servicios públicos, sino también la integridad y fiabilidad de las decisiones adoptadas mediante estos sistemas.

2 Veale, M., Borgesius, F. Z., «Demystifying the Draft EU Artificial Intelligence Act», *Computer Law Review International*, 2021.

3 Unión Europea, *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).*

4 Unión Europea, *Directiva (UE) 2022/2555 del Parlamento Europeo y del Consejo de 14 de diciembre de 2022 relativa a las medidas destinadas a garantizar un elevado nivel común de ciberseguridad en toda la Unión, por la que se modifican el Reglamento (UE) n.o 910/2014 y la Directiva (UE) 2018/1972 y por la que se deroga la Directiva (UE) 2016/1148 (Directiva SRI 2).*

5 Kuner, C., Bygrave, L. A., Docksey, C. (Eds.), *The EU General Data Protection Regulation (GDPR): A Commentary – 2021 Update*, Maastricht Faculty of Law Working Paper, 2021.

Junto a estos instrumentos vinculantes, el marco normativo europeo se completa con una serie de iniciativas de *soft law*, entre las que destacan las Directrices éticas para una inteligencia artificial fiable elaboradas por el Grupo de Expertos de Alto Nivel en Inteligencia Artificial de la Comisión Europea, así como diversas comunicaciones y estrategias en materia de inteligencia artificial. Estos instrumentos, si bien carecen de fuerza jurídica vinculante, desempeñan un papel relevante en la configuración de estándares interpretativos y en la orientación de las políticas públicas en la materia[6].

Desde una perspectiva comparada, la situación en América Latina presenta un grado de desarrollo normativo notablemente desigual. Si bien algunos países han comenzado a adoptar estrategias nacionales de inteligencia artificial y marcos regulatorios incipientes, la mayoría carece aún de una regulación específica equivalente al modelo europeo. En este contexto, la gobernanza de la inteligencia artificial se articula fundamentalmente a través de normas generales particularmente en materia de protección de datos y de iniciativas programáticas que no siempre se traducen en mecanismos efectivos de control y supervisión.

Esta asimetría regulatoria pone de manifiesto la existencia de una brecha significativa entre la Unión Europea y América Latina en términos de institucionalización de la gobernanza algorítmica. Mientras que en el contexto europeo se observa una tendencia hacia la densificación normativa y la formalización de obligaciones jurídicas, en el ámbito latinoamericano predomina un enfoque más flexible, pero también más fragmentado, lo que puede incrementar los riesgos asociados a la utilización de sistemas de inteligencia artificial en el sector público.

En definitiva, el análisis del marco normativo revela que, pese a los avances significativos en el plano regulatorio, persisten importantes desafíos en términos de coherencia, coordinación e implementación efectiva. La coexistencia de múltiples instrumentos normativos, con distintos niveles de obligatoriedad y ámbitos de aplicación, exige una interpretación sistemática que permita articular un modelo de gobernanza capaz de garantizar el equilibrio entre innovación tecnológica, eficiencia administrativa y protección de los derechos fundamentales.

III. Desafíos de la implementación de la inteligencia artificial en la Administración Pública

1. La opacidad algorítmica y los límites de la explicabilidad

Uno de los principales desafíos derivados de la implementación de sistemas de inteligencia artificial en la Administración Pública se encuentra vinculado a la opacidad algorítmica y a las limitaciones existentes en términos de explicabilidad de las decisiones automatizadas. La progresiva utilización de sistemas basados en aprendizaje automático introduce importantes dificultades para comprender el funcionamiento interno de los modelos algorítmicos, especialmente

6 Comisión Europea, «*Ethics Guidelines for Trustworthy AI»,* High-Level Expert Group on Artificial Intelligence, 2019.

en aquellos supuestos en los que las decisiones administrativas se adoptan a partir de procesos complejos de tratamiento masivo de datos.

En este contexto, la doctrina especializada ha advertido que los sistemas de inteligencia artificial, particularmente aquellos sustentados en técnicas de machine learning, presentan limitaciones estructurales que dificultan la reconstrucción lógica del proceso decisional. Como señalan Wachter, Mittelstadt y Floridi, «the right to explanation is not a general right under the GDPR, but a limited safeguard applicable in specific contexts»[7], lo que pone de manifiesto que el Reglamento General de Protección de Datos no reconoce un derecho absoluto a obtener una explicación completa del funcionamiento algorítmico, sino únicamente determinadas garantías vinculadas a decisiones automatizadas concretas.

La problemática adquiere una especial relevancia en el ámbito de la Administración Pública, donde la utilización de sistemas automatizados no afecta exclusivamente a cuestiones organizativas internas, sino que puede incidir directamente en el ejercicio de derechos fundamentales y en la adopción de decisiones con efectos jurídicos sobre la ciudadanía. La opacidad algorítmica dificulta la motivación adecuada de los actos administrativos, limita la posibilidad de control jurisdiccional y reduce las capacidades efectivas de impugnación por parte de los ciudadanos afectados.

Desde esta perspectiva, la utilización de algoritmos opacos tensiona los principios clásicos del Derecho administrativo, particularmente los principios de transparencia, motivación y sometimiento pleno a la ley y al Derecho. Como advierte Pasquale, «black box algorithms challenge the very foundations of accountability in public administration»[8], al introducir mecanismos decisionales cuyo funcionamiento interno resulta inaccesible incluso para las propias entidades que los utilizan.

En el ámbito doctrinal español, Agustí Cerrillo i Martínez ha puesto de relieve que la utilización de algoritmos por parte de las Administraciones Públicas exige reforzar las exigencias de transparencia y trazabilidad de las decisiones automatizadas, especialmente cuando estas producen efectos jurídicos relevantes sobre los ciudadanos. En este sentido, el autor advierte que la automatización administrativa no puede traducirse en una reducción de las garantías propias del procedimiento administrativo ni en una disminución de las obligaciones de motivación y control que corresponden a los poderes públicos.

La dificultad de acceso al funcionamiento interno de los sistemas algorítmicos plantea, además, importantes problemas en relación con el principio de explicabilidad. Aunque el Reglamento de Inteligencia Artificial incorpora

7 Wachter, S., Mittelstadt, B., Floridi, L., «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation», *International Data Privacy Law*, vol. 7, n.º 2 (2017), pp. 76-99. Traducción propia: «el derecho a la explicación no es un derecho general en el marco del RGPD, sino una garantía limitada aplicable en contextos específicos».

8 Pasquale, F., *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015.

obligaciones reforzadas de transparencia para determinados sistemas de alto riesgo, la aplicación práctica de estas exigencias continúa enfrentándose a obstáculos derivados de la complejidad técnica de los modelos y de la existencia de sistemas cuya lógica resulta difícilmente interpretable incluso para sus propios desarrolladores.

A ello se añade que la opacidad algorítmica no constituye únicamente un problema técnico, sino también institucional y democrático. La imposibilidad de comprender adecuadamente cómo se adoptan determinadas decisiones automatizadas puede erosionar la confianza de los ciudadanos en las instituciones públicas y debilitar los mecanismos tradicionales de control de la actividad administrativa. Como ha señalado la doctrina, la transparencia algorítmica no puede limitarse a la mera divulgación formal de información técnica, sino que debe garantizar una comprensión efectiva del funcionamiento y de los criterios utilizados por el sistema en la adopción de decisiones.

En consecuencia, puede afirmarse que la opacidad algorítmica representa uno de los principales límites para la implementación efectiva de sistemas de inteligencia artificial compatibles con los principios del Estado de Derecho. La incorporación de estas tecnologías en la Administración Pública exige avanzar hacia modelos de gobernanza que permitan garantizar niveles suficientes de transparencia, explicabilidad y control jurídico, evitando que la automatización de decisiones administrativas se convierta en un espacio inmune a la supervisión democrática y jurisdiccional.

2. La fragmentación de la responsabilidad en decisiones automatizadas

Otro de los desafíos centrales derivados de la incorporación de sistemas de inteligencia artificial en la Administración Pública se encuentra relacionado con la compleja atribución de responsabilidad jurídica en contextos de decisión automatizada. A diferencia de los modelos administrativos tradicionales, en los que la actuación puede imputarse de manera relativamente clara a un órgano administrativo concreto, los sistemas algorítmicos introducen dinámicas de intervención múltiple que dificultan la identificación precisa de los sujetos responsables.

En efecto, el diseño, desarrollo, implementación y funcionamiento de los sistemas de inteligencia artificial suele involucrar a una pluralidad de actores, entre los que se encuentran desarrolladores de software, empresas proveedoras de servicios tecnológicos, entidades encargadas del tratamiento de datos y las propias Administraciones Públicas que finalmente utilizan estos sistemas para la adopción de decisiones administrativas. Esta estructura fragmentada genera importantes dificultades para determinar el alcance de la responsabilidad jurídica en caso de error, discriminación algorítmica o vulneración de derechos fundamentales.

La doctrina especializada ha advertido que esta distribución funcional de tareas puede derivar en auténticos vacíos de responsabilidad. En este sentido, Bovens señala que «the diffusion of responsibility in algorithmic systems risks

creating accountability gaps»[9], poniendo de manifiesto que la complejidad técnica de los sistemas automatizados puede diluir la capacidad efectiva de identificar responsables concretos dentro de la cadena decisional.

La problemática resulta especialmente relevante en el ámbito público, donde las decisiones automatizadas no se limitan a operaciones técnicas neutras, sino que se integran directamente en el ejercicio del poder administrativo. La utilización de sistemas algorítmicos en procedimientos relacionados con prestaciones sociales, selección de beneficiarios, gestión tributaria o evaluación de riesgos administrativos puede producir efectos jurídicos significativos sobre los ciudadanos, lo que exige garantizar mecanismos efectivos de control y responsabilidad.

Desde esta perspectiva, la automatización de la actividad administrativa no elimina la responsabilidad de la Administración Pública, sino que obliga a reinterpretar sus fundamentos tradicionales. La decisión automatizada continúa formando parte de la actuación administrativa y, en consecuencia, debe permanecer sometida a los principios de legalidad, control jurisdiccional y responsabilidad patrimonial propios del Derecho administrativo. La complejidad tecnológica del sistema no puede convertirse en un mecanismo de exclusión o debilitamiento de las garantías jurídicas de los ciudadanos.

En el ámbito doctrinal español, Eduardo Gamero Casado[10] ha advertido sobre la necesidad de adaptar las categorías clásicas del Derecho administrativo a los nuevos escenarios de automatización pública, señalando que la transformación digital de las Administraciones exige reforzar los mecanismos de control jurídico sobre las decisiones automatizadas. En particular, la utilización de inteligencia artificial en el sector público obliga a replantear cuestiones vinculadas a la motivación de los actos administrativos, la trazabilidad de las decisiones y la posibilidad de revisión efectiva por parte de los órganos jurisdiccionales.

A ello se añade que muchos sistemas de inteligencia artificial funcionan mediante procesos de aprendizaje continuo, circunstancia que incrementa la dificultad para reconstruir el razonamiento seguido por el sistema en cada decisión concreta. Como señalan Burr y Cramer, «many machine-learning systems are effectively black boxes whose internal logic is difficult or impossible to interpret»[11], lo que limita significativamente la capacidad de identificar relaciones causales claras entre el funcionamiento del algoritmo y el daño eventualmente producido.

La fragmentación de la responsabilidad también plantea importantes problemas desde la perspectiva democrática e institucional. La dificultad para

9 Bovens, M., «Analysing and Assessing Accountability: A Conceptual Framework», *European Law Journal*, 2007. Traducción propia: «La difusión de la responsabilidad en los sistemas algorítmicos corre el riesgo de generar vacíos de rendición de cuentas».

10 Gamero Casado, E., «Necesidad de motivación e invalidez de los actos administrativos sustentados en inteligencia artificial o en algoritmos», *Almacén de Derecho*, 2021.

11 Burr, C., Cramer, H., *The EU's Artificial Intelligence Act: Risks and Regulatory Challenges*, 2018, p. 3. Traducción propia: «Muchos sistemas de aprendizaje automático son, en la práctica, cajas negras cuya lógica interna resulta difícil o incluso imposible de interpretar».

identificar responsables concretos puede erosionar los principios de *accountability* y rendición de cuentas que caracterizan al Estado de Derecho, especialmente cuando las decisiones automatizadas producen consecuencias relevantes sobre derechos individuales. En este contexto, la existencia de sistemas complejos y técnicamente opacos corre el riesgo de generar espacios de actuación administrativa insuficientemente sometidos a mecanismos efectivos de supervisión pública y control jurídico.

Por ello, la gobernanza algorítmica exige avanzar hacia modelos de responsabilidad compartida y supervisión continua que permitan garantizar la trazabilidad de las decisiones automatizadas y la identificación clara de obligaciones jurídicas para todos los actores que intervienen en el ciclo de vida del sistema. La construcción de estos mecanismos resulta esencial para evitar que la complejidad tecnológica termine debilitando las garantías tradicionales de responsabilidad administrativa y protección de los derechos fundamentales.

3. La supervisión humana significativa y los límites del control humano

La supervisión humana constituye uno de los principios centrales sobre los que se articula el modelo europeo de regulación de la inteligencia artificial, especialmente en relación con los sistemas considerados de alto riesgo. Tanto el Reglamento de Inteligencia Artificial como diversos instrumentos de *soft law* elaborados por las instituciones europeas insisten en la necesidad de garantizar una intervención humana efectiva sobre los procesos automatizados de toma de decisiones. Sin embargo, la implementación práctica de esta exigencia continúa planteando importantes dificultades jurídicas, técnicas y organizativas.

En efecto, la mera existencia formal de intervención humana no garantiza por sí misma un control real sobre el funcionamiento de los sistemas algorítmicos. En numerosos supuestos, la complejidad técnica de los modelos de inteligencia artificial y la dificultad para interpretar sus resultados convierten la supervisión humana en un mecanismo limitado o incluso meramente simbólico. Como ha señalado la Comisión Europea, «human oversight should ensure that AI systems do not undermine human autonomy or cause other adverse effects»[12], lo que implica que la intervención humana debe poseer capacidad material suficiente para comprender, evaluar y, en su caso, corregir las decisiones automatizadas.

La problemática adquiere una especial relevancia en el ámbito de la Administración Pública, donde las decisiones automatizadas pueden proyectarse sobre derechos fundamentales y sobre situaciones jurídicas individuales de los ciudadanos. La utilización de sistemas de inteligencia artificial en procedimientos administrativos exige garantizar que los operadores públicos mantengan un grado suficiente de control sobre el funcionamiento del sistema y sobre las decisiones finalmente adoptadas.

12 Comisión Europea, *Ethics Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence*, 2019, p. 17. Traducción propia: «La supervisión humana debe garantizar que los sistemas de IA no menoscaben la autonomía humana ni provoquen otros efectos adversos».

No obstante, en la práctica, la supervisión humana se enfrenta a importantes limitaciones derivadas de la propia configuración tecnológica de muchos sistemas algorítmicos. La creciente sofisticación de los modelos de *machine learning* dificulta que los operadores públicos puedan comprender plenamente los criterios utilizados por el sistema o detectar posibles errores, sesgos o disfunciones en el proceso decisional. En este contexto, existe el riesgo de que la intervención humana termine reduciéndose a una validación automática de los resultados generados por el algoritmo, sin capacidad efectiva de revisión crítica.

Desde la doctrina española, José Luis Piñar Mañas ha subrayado que la utilización de inteligencia artificial en el sector público exige reforzar las garantías vinculadas a la protección de datos, la transparencia y el control de las decisiones automatizadas, especialmente cuando estas afectan al ejercicio de derechos por parte de los ciudadanos. En este sentido, la supervisión humana no puede interpretarse como un requisito puramente formal, sino como una garantía material vinculada a la preservación de la autonomía decisional y al sometimiento pleno de la Administración al ordenamiento jurídico.

Asimismo, la doctrina ha advertido que la automatización excesiva puede generar fenómenos de dependencia tecnológica dentro de las Administraciones Públicas. La progresiva confianza en sistemas automatizados de decisión puede provocar una disminución de la capacidad crítica de los operadores humanos y una tendencia a aceptar las recomendaciones algorítmicas como objetivas o técnicamente infalibles. Este fenómeno, conocido en algunos sectores doctrinales como *automation bias*, incrementa el riesgo de consolidar decisiones erróneas o discriminatorias bajo una apariencia de neutralidad tecnológica.

A ello se añade la insuficiente preparación técnica y organizativa de muchas Administraciones Públicas para asumir un control efectivo sobre sistemas complejos de inteligencia artificial. La supervisión humana significativa exige no solo la existencia de intervención humana, sino también la formación especializada del personal, la disponibilidad de recursos técnicos adecuados y la existencia de protocolos claros de revisión y auditoría de decisiones automatizadas.

Desde esta perspectiva, la supervisión humana debe concebirse como un elemento estructural de la gobernanza algorítmica y no como una garantía meramente declarativa. La intervención humana únicamente resulta efectiva cuando existe una capacidad real de comprender el funcionamiento del sistema, cuestionar sus resultados y adoptar decisiones alternativas en aquellos supuestos en los que el algoritmo produzca resultados incompatibles con los principios de legalidad, proporcionalidad o igualdad.

En consecuencia, puede afirmarse que uno de los principales desafíos de la implementación de inteligencia artificial en la Administración Pública consiste en garantizar formas de supervisión humana que sean materialmente efectivas y jurídicamente relevantes. La legitimidad de las decisiones automatizadas depende, en gran medida, de que las Administraciones conserven la capacidad real de controlar el funcionamiento de los sistemas algorítmicos y de evitar que la automatización termine desplazando las garantías tradicionales del derecho administrativo.

4. Capacidades institucionales, cultura administrativa y adaptación organizativa

La implementación efectiva de sistemas de inteligencia artificial en la Administración Pública no depende exclusivamente de la existencia de marcos normativos adecuados, sino también de la capacidad institucional de las organizaciones públicas para integrar estas tecnologías dentro de sus estructuras administrativas y operativas. En este sentido, uno de los principales desafíos radica en la insuficiente preparación técnica, organizativa y cultural de muchas Administraciones para asumir procesos complejos de automatización decisional.

La incorporación de inteligencia artificial en el sector público implica transformaciones que trascienden la mera adquisición de herramientas tecnológicas. La utilización de sistemas algorítmicos exige adaptar procedimientos administrativos, redefinir dinámicas organizativas y desarrollar nuevas capacidades institucionales orientadas a garantizar un uso adecuado, seguro y jurídicamente compatible de estas tecnologías. Sin embargo, en numerosos casos, la digitalización administrativa continúa desarrollándose sobre estructuras organizativas diseñadas para modelos tradicionales de gestión pública, lo que genera tensiones entre innovación tecnológica y funcionamiento institucional.

Desde esta perspectiva, uno de los principales riesgos consiste en asumir que la automatización tecnológica produce automáticamente mayores niveles de eficiencia administrativa. La literatura especializada ha advertido que la incorporación de sistemas de inteligencia artificial sin una revisión previa de los procesos administrativos puede generar nuevas formas de ineficiencia, aumentar la complejidad organizativa y dificultar los mecanismos de control interno.

Asimismo, la implementación de inteligencia artificial en la Administración Pública requiere recursos técnicos y humanos altamente especializados. La supervisión efectiva de sistemas algorítmicos exige personal con formación interdisciplinar capaz de comprender aspectos jurídicos, tecnológicos y organizativos vinculados al funcionamiento de la inteligencia artificial. No obstante, muchas Administraciones continúan presentando importantes limitaciones en términos de capacitación técnica, lo que dificulta el desarrollo de mecanismos efectivos de supervisión y auditoría.

En este contexto, la formación de los operadores públicos adquiere una relevancia central dentro de los modelos de gobernanza algorítmica. La utilización de sistemas automatizados exige que los empleados públicos dispongan de capacidades suficientes para interpretar resultados algorítmicos, detectar posibles riesgos y comprender las implicaciones jurídicas derivadas de la utilización de inteligencia artificial en procedimientos administrativos.

La doctrina también ha puesto de relieve que la transformación digital de las Administraciones Públicas requiere una evolución de la cultura organizativa institucional. La incorporación de tecnologías disruptivas como la inteligencia artificial obliga a superar modelos administrativos rígidos y avanzar hacia estructuras más flexibles, interoperables y adaptativas. En este sentido, la resistencia institucional al cambio, la fragmentación administrativa y la falta de

coordinación entre departamentos pueden limitar significativamente la efectividad de las estrategias de digitalización pública.

Desde el ámbito doctrinal español, Agustí Cerrillo i Martínez ha señalado que la gobernanza de la inteligencia artificial en el sector público exige reforzar las capacidades institucionales de las Administraciones y desarrollar mecanismos organizativos que permitan garantizar la trazabilidad, supervisión y control de los sistemas automatizados. La transformación digital administrativa no puede reducirse a una mera cuestión tecnológica, sino que requiere una adaptación estructural de las instituciones públicas.

A ello se añade la necesidad de integrar mecanismos de *compliance* digital dentro de las estructuras administrativas. La utilización de sistemas de inteligencia artificial exige la creación de protocolos internos de evaluación de riesgos, supervisión continua y auditoría tecnológica que permitan garantizar el cumplimiento de principios jurídicos como la transparencia, proporcionalidad, igualdad y protección de datos personales.

En consecuencia, puede afirmarse que uno de los principales desafíos de la implementación de inteligencia artificial en la Administración Pública reside en la capacidad institucional para gestionar adecuadamente estas tecnologías. La efectividad de los modelos de gobernanza algorítmica depende no solo del desarrollo normativo, sino también de la existencia de estructuras organizativas, capacidades técnicas y culturas administrativas compatibles con las exigencias derivadas de la automatización del poder público.

5. América Latina y la brecha de gobernanza algorítmica

Los desafíos asociados a la implementación de sistemas de inteligencia artificial en la Administración Pública adquieren una dimensión especialmente compleja en el contexto latinoamericano, donde la incorporación de tecnologías automatizadas avanza de manera desigual y, en muchos casos, sin marcos regulatorios consolidados ni estructuras institucionales suficientemente desarrolladas para garantizar un control efectivo sobre los sistemas algorítmicos.

A diferencia del modelo europeo, caracterizado por una progresiva densificación normativa en torno a la inteligencia artificial, la protección de datos y la ciberseguridad, en América Latina predomina un escenario fragmentado, marcado por estrategias nacionales heterogéneas y por la ausencia de mecanismos integrales de gobernanza algorítmica. Si bien algunos países han comenzado a desarrollar políticas públicas orientadas a fomentar el uso de inteligencia artificial en el sector público, la implementación práctica de estas iniciativas continúa enfrentándose a importantes limitaciones institucionales, técnicas y regulatorias.

En este contexto, la utilización de sistemas automatizados en la Administración Pública puede generar riesgos significativos para la protección de los derechos fundamentales, especialmente en aquellos entornos donde las capacidades institucionales de supervisión y control resultan limitadas. La ausencia de estándares homogéneos de transparencia, auditoría y explicabilidad

incrementa la vulnerabilidad de los ciudadanos frente a decisiones automatizadas potencialmente arbitrarias o discriminatorias.

Asimismo, la debilidad de los mecanismos de control administrativo y jurisdiccional dificulta la posibilidad de cuestionar eficazmente decisiones adoptadas mediante sistemas algorítmicos. La falta de regulación específica sobre inteligencia artificial en numerosos ordenamientos latinoamericanos provoca que muchas decisiones automatizadas se desarrollen en espacios normativos ambiguos, donde las garantías tradicionales del procedimiento administrativo no siempre resultan claramente aplicables.

Diversos informes e investigaciones regionales han advertido sobre la necesidad de construir modelos de gobernanza digital capaces de equilibrar innovación tecnológica y protección efectiva de derechos fundamentales, particularmente en contextos de creciente automatización de la actividad pública[13]. En este sentido, la transformación digital del Estado no puede desarrollarse al margen de principios como la transparencia, la responsabilidad y la protección de datos personales.

A ello se añade que gran parte de los países latinoamericanos presentan importantes asimetrías en términos de capacidades tecnológicas e institucionales, lo que dificulta la implementación de mecanismos avanzados de auditoría y supervisión algorítmica. La dependencia de proveedores tecnológicos externos y la limitada disponibilidad de personal especializado incrementan el riesgo de generar formas de automatización insuficientemente controladas por las propias Administraciones Públicas.

La problemática también se proyecta sobre el ámbito democrático e institucional. La utilización de inteligencia artificial en contextos caracterizados por elevados niveles de desigualdad social y brechas digitales puede profundizar situaciones de exclusión y discriminación estructural, especialmente cuando los sistemas automatizados son entrenados con datos sesgados o implementados sin mecanismos adecuados de evaluación de impacto.

Desde esta perspectiva, la experiencia europea pone de manifiesto la importancia de construir modelos regulatorios preventivos y estructurados que permitan anticipar riesgos antes de la implementación efectiva de sistemas automatizados. No obstante, la eventual recepción de estos modelos en América Latina exige tener en consideración las particularidades institucionales, económicas y administrativas propias de la región, evitando procesos de trasplante normativo descontextualizados.

En consecuencia, puede afirmarse que la gobernanza algorítmica en América Latina continúa enfrentándose a importantes desafíos relacionados con la debilidad institucional, la fragmentación regulatoria y la insuficiente consolidación de mecanismos de supervisión tecnológica. La construcción de modelos efectivos de inteligencia artificial pública en la región requiere no solo avances normativos, sino también el fortalecimiento de capacidades institucionales y el desarrollo de

13 OECD *et al.*, *Latin American Economic Outlook 2023: Investing in Sustainable Development*, OECD Publishing, 2023.

estructuras de control compatibles con los principios del Estado de Derecho y la protección de los derechos fundamentales.

Ello evidencia que la implementación de sistemas de inteligencia artificial en el ámbito público no constituye únicamente un desafío tecnológico, sino también un problema de gobernanza democrática, legitimidad institucional y preservación efectiva de las garantías propias del Estado de Derecho.

IV. Responsabilidad y control en los sistemas algorítmicos

1. La transformación de los modelos tradicionales de responsabilidad administrativa

La incorporación de sistemas de inteligencia artificial en la Administración Pública plantea la necesidad de revisar los fundamentos tradicionales de la responsabilidad jurídica, en la medida en que los esquemas clásicos de imputación resultan insuficientes para dar respuesta a las nuevas formas de producción de decisiones automatizadas. En este contexto, la interacción entre tecnología y Derecho introduce una complejidad adicional derivada de la pluralidad de actores que intervienen en el diseño, desarrollo e implementación de los sistemas algorítmicos.

En efecto, la configuración de estos sistemas implica la participación de distintos sujetos, entre los que se encuentran desarrolladores de software, proveedores tecnológicos, entidades encargadas del tratamiento de datos y las propias Administraciones Públicas que implementan y utilizan sistemas automatizados en el ejercicio de funciones públicas. Esta fragmentación funcional dificulta la identificación de un sujeto claramente responsable en caso de error, daño o vulneración de derechos fundamentales, generando tensiones con los modelos tradicionales de responsabilidad administrativa basados en la imputación directa de la actuación a un órgano público concreto.

La problemática adquiere una especial relevancia en aquellos supuestos en los que las decisiones automatizadas producen efectos jurídicos sobre los ciudadanos, particularmente en ámbitos vinculados a prestaciones sociales, procedimientos sancionadores, gestión tributaria o selección automatizada de beneficiarios de servicios públicos. En estos contextos, la automatización de la actividad administrativa no puede traducirse en una disminución de las garantías jurídicas inherentes al ejercicio del poder público[14].

Como ha señalado la doctrina, la complejidad tecnológica y la transparencia limitada de muchos sistemas de inteligencia artificial no permiten garantizar por sí solas mecanismos efectivos de control, en la medida en que «transparency alone does not necessarily produce accountability»[15]. La mera existencia de información sobre el funcionamiento del sistema no asegura necesariamente la

14 Sunstein, C. R., *Algorithms, Correcting Biases, and Public Policy*, Harvard Law School, 2018.

15 Ananny, M., Crawford, K., «Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability», *New Media & Society*, vol. 20, n.º

posibilidad de exigir responsabilidad jurídica ni garantiza una supervisión efectiva de las decisiones automatizadas.

Desde esta perspectiva, la utilización de inteligencia artificial en el ámbito público no elimina la responsabilidad de la Administración, sino que exige reinterpretar sus fundamentos tradicionales a la luz de los nuevos entornos tecnológicos. La decisión automatizada continúa integrándose en el ejercicio de la función administrativa y, en consecuencia, debe permanecer sometida a los principios de legalidad, responsabilidad patrimonial y control jurisdiccional propios del Derecho administrativo.

En este sentido, y como ya se advirtió al analizar los problemas derivados de la fragmentación de responsabilidad en los sistemas automatizados, la doctrina administrativista española ha señalado la necesidad de adaptar las categorías clásicas del Derecho administrativo a los nuevos escenarios de automatización pública, particularmente en relación con la motivación de los actos administrativos sustentados en algoritmos y sistemas de inteligencia artificial

Asimismo, la creciente automatización de decisiones administrativas obliga a replantear el alcance de la responsabilidad patrimonial de las Administraciones Públicas en contextos de funcionamiento algorítmico. La dificultad de identificar con precisión el origen del error o del daño producido por un sistema automatizado puede generar situaciones de incertidumbre jurídica que afecten negativamente a la protección efectiva de los ciudadanos frente a actuaciones administrativas automatizadas.

La doctrina también ha advertido que la distribución de responsabilidades entre múltiples operadores tecnológicos puede favorecer fenómenos de dilución de responsabilidad incompatibles con los principios del Estado de Derecho. En consecuencia, resulta necesario avanzar hacia modelos de responsabilidad capaces de garantizar mecanismos efectivos de imputación y control, evitando que la complejidad técnica de los sistemas automatizados termine debilitando las garantías tradicionales del Derecho administrativo.

Por ello, puede afirmarse que la incorporación de inteligencia artificial en la Administración Pública exige una evolución de los modelos tradicionales de responsabilidad jurídica hacia esquemas más complejos y adaptativos, capaces de responder adecuadamente a la naturaleza distribuida y dinámica de los sistemas algorítmicos sin renunciar a los principios estructurales de responsabilidad pública y protección de los derechos fundamentales.

2. Opacidad algorítmica y dificultades de imputación causal

No obstante, la atribución de responsabilidad jurídica en contextos de decisión automatizada se ve especialmente condicionada por la opacidad y la autonomía relativa de los sistemas algorítmicos. La dificultad para reconstruir el proceso decisional y para identificar los factores que han influido en el resultado

3 (2018), p. 979. Traducción propia: «La transparencia por sí sola no produce necesariamente rendición de cuentas».

final limita la posibilidad de establecer una relación causal clara, lo que puede debilitar los mecanismos tradicionales de control jurídico.

La problemática resulta particularmente intensa en aquellos sistemas basados en técnicas de aprendizaje automático, donde los modelos algorítmicos operan mediante procesos complejos de tratamiento de datos cuya lógica interna puede resultar difícilmente interpretable incluso para los propios desarrolladores del sistema. La doctrina especializada ha advertido que muchos sistemas basados en *machine learning* funcionan mediante dinámicas de razonamiento automatizado cuya reconstrucción jurídica y técnica presenta importantes limitaciones, circunstancia que dificulta la comprensión efectiva de los criterios utilizados en la adopción de decisiones automatizadas.

En el ámbito de la Administración Pública, esta opacidad algorítmica plantea importantes tensiones respecto de principios esenciales del Derecho administrativo, particularmente en relación con la motivación de los actos administrativos, la transparencia de la actuación pública y la posibilidad de control jurisdiccional efectivo. La dificultad para conocer los criterios utilizados por el sistema automatizado puede limitar significativamente la capacidad de los ciudadanos para impugnar decisiones administrativas sustentadas en inteligencia artificial.

Asimismo, la opacidad de los sistemas algorítmicos dificulta la determinación precisa del origen del daño o de la eventual irregularidad producida durante el funcionamiento del sistema. La intervención simultánea de múltiples operadores tecnológicos, junto con la complejidad técnica de los modelos automatizados, genera importantes obstáculos para establecer relaciones causales claras entre el comportamiento del algoritmo y las consecuencias jurídicas derivadas de su utilización.

A ello se añade que determinados sistemas de inteligencia artificial presentan capacidades de aprendizaje continuo y adaptación dinámica, lo que incrementa aún más las dificultades de supervisión y control jurídico. La evolución constante de los modelos algorítmicos puede alterar los criterios de decisión inicialmente programados, dificultando la trazabilidad del proceso automatizado y limitando la capacidad de anticipar determinados riesgos asociados a su funcionamiento.

Desde esta perspectiva, la opacidad algorítmica no constituye únicamente un problema técnico, sino también un desafío jurídico e institucional que afecta directamente a la legitimidad de las decisiones automatizadas adoptadas por las Administraciones Públicas. La imposibilidad de comprender adecuadamente cómo se produce una decisión administrativa automatizada puede erosionar la confianza ciudadana en las instituciones públicas y debilitar las garantías propias del Estado de Derecho.

Como ya se advirtió en relación con los problemas de explicabilidad analizados en el apartado anterior, la mera existencia de mecanismos formales de transparencia no garantiza necesariamente una comprensión efectiva del funcionamiento de los sistemas algorítmicos. La complejidad técnica de muchos modelos automatizados exige avanzar hacia mecanismos reforzados de trazabilidad,

auditoría y supervisión continua que permitan garantizar un control real sobre la actuación administrativa automatizada.

En consecuencia, puede afirmarse que las dificultades de imputación causal derivadas de la opacidad algorítmica representan uno de los principales obstáculos para la construcción de modelos efectivos de responsabilidad jurídica en contextos de inteligencia artificial pública. La protección adecuada de los derechos fundamentales exige desarrollar mecanismos capaces de garantizar no solo la transparencia formal de los sistemas automatizados, sino también la posibilidad efectiva de reconstruir, revisar y cuestionar las decisiones adoptadas mediante inteligencia artificial.

3. Auditorías algorítmicas y mecanismos preventivos de control

La creciente utilización de sistemas de inteligencia artificial en la Administración Pública ha puesto de manifiesto la insuficiencia de los modelos tradicionales de control jurídico basados exclusivamente en mecanismos de revisión ex post. La capacidad de los sistemas automatizados para adoptar decisiones a gran escala y de manera prácticamente instantánea exige desarrollar formas de supervisión preventiva que permitan identificar riesgos antes de que estos produzcan efectos lesivos sobre los derechos de los ciudadanos.

En este contexto, la gobernanza algorítmica[16] requiere incorporar mecanismos de control *ex ante* orientados a evaluar la legalidad, proporcionalidad y fiabilidad de los sistemas de inteligencia artificial antes de su implementación efectiva en procedimientos administrativos. La supervisión posterior de las decisiones automatizadas resulta insuficiente cuando los sistemas algorítmicos operan de forma continuada y afectan simultáneamente a un elevado número de personas.

Desde esta perspectiva, adquieren una relevancia central las auditorías algorítmicas, entendidas como instrumentos destinados a verificar el cumplimiento de requisitos jurídicos, técnicos y éticos durante el diseño, desarrollo e implementación de sistemas de inteligencia artificial. Estas auditorías permiten identificar riesgos asociados a sesgos algorítmicos[17], deficiencias de transparencia, problemas de trazabilidad o posibles vulneraciones de derechos fundamentales derivadas del funcionamiento del sistema automatizado.

La doctrina especializada ha señalado que las auditorías algorítmicas constituyen una herramienta esencial para reforzar la confianza institucional en el uso de inteligencia artificial dentro del sector público, particularmente en aquellos supuestos en los que las decisiones automatizadas producen efectos jurídicos relevantes sobre los ciudadanos. En este sentido, los mecanismos de auditoría no deben limitarse únicamente a aspectos técnicos vinculados al funcionamiento

16 Yeung, K., «Algorithmic regulation: A critical interrogation», *Regulation & Governance*, vol. 12, n.º 4 (2018), pp. 505-523.

17 Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., Yu, H., «Accountable Algorithms», *University of Pennsylvania Law Review*, vol. 165, 2017, pp. 633-705.

del software, sino extenderse también a la evaluación de impactos jurídicos, organizativos y éticos asociados al sistema.

Asimismo, la implementación de mecanismos preventivos de control exige reforzar las obligaciones de documentación y trazabilidad de los sistemas automatizados. La existencia de registros adecuados sobre el funcionamiento del algoritmo, las fuentes de datos utilizadas y los criterios aplicados en la toma de decisiones constituye un elemento esencial para garantizar la posibilidad de revisión posterior y para facilitar la exigencia de responsabilidad jurídica en caso de funcionamiento irregular.

A ello se añade la necesidad de desarrollar protocolos institucionales de evaluación continua que permitan supervisar el comportamiento dinámico de los sistemas algorítmicos una vez implementados. La naturaleza evolutiva de determinados modelos de inteligencia artificial exige mecanismos permanentes de seguimiento orientados a detectar modificaciones inesperadas en los criterios de decisión o la aparición de riesgos no previstos inicialmente.

En el ámbito europeo, el Reglamento de Inteligencia Artificial avanza precisamente hacia modelos de supervisión preventiva basados en evaluaciones de riesgo, obligaciones reforzadas de transparencia y exigencias de control continuo sobre determinados sistemas considerados de alto riesgo. Este enfoque refleja una transformación progresiva de los modelos clásicos de control administrativo hacia esquemas más preventivos y adaptativos, compatibles con la complejidad técnica de los sistemas automatizados.

Desde una perspectiva más amplia, las auditorías algorítmicas deben integrarse dentro de modelos de *compliance* digital orientados a garantizar un uso responsable de la inteligencia artificial en el ámbito público. La implementación de protocolos internos de supervisión, evaluación de impacto y gestión de riesgos constituye un elemento esencial para asegurar que la automatización administrativa permanezca sometida a los principios de legalidad, transparencia y protección de los derechos fundamentales.

En consecuencia, puede afirmarse que la consolidación de mecanismos preventivos de control representa una condición indispensable para el desarrollo de modelos legítimos de inteligencia artificial pública. La complejidad y capacidad expansiva de los sistemas algorítmicos exige superar enfoques tradicionales de supervisión reactiva y avanzar hacia estructuras permanentes de auditoría, evaluación y gobernanza continua de las decisiones automatizadas.

4. Supervisión humana y *accountability* en la gobernanza algorítmica

La creciente automatización de decisiones administrativas exige reforzar los mecanismos de supervisión humana dentro de los modelos de gobernanza algorítmica. La utilización de sistemas de inteligencia artificial en el ámbito público no puede conducir a la sustitución completa de la intervención humana en aquellos procedimientos que afectan al ejercicio de derechos o producen efectos jurídicos relevantes sobre los ciudadanos. En consecuencia, la supervisión

humana constituye una garantía esencial para preservar la legitimidad y el control democrático de la actuación administrativa automatizada.

No obstante, la exigencia de supervisión humana no debe interpretarse en un sentido meramente formal. La simple existencia nominal de intervención humana resulta insuficiente cuando los operadores públicos carecen de capacidad real para comprender, evaluar o corregir las decisiones adoptadas por los sistemas algorítmicos. En este contexto, la supervisión efectiva exige que los responsables administrativos dispongan de conocimientos técnicos y jurídicos suficientes para ejercer un control material sobre el funcionamiento del sistema automatizado.

Como ya se advirtió en relación con las limitaciones derivadas de la opacidad algorítmica, muchos sistemas de inteligencia artificial operan mediante dinámicas técnicas de elevada complejidad que dificultan la comprensión efectiva de los criterios utilizados en la toma de decisiones. Esta circunstancia incrementa el riesgo de que la intervención humana se reduzca a una validación automática de los resultados generados por el algoritmo, sin una revisión crítica real de sus consecuencias jurídicas o de sus posibles efectos discriminatorios.

A ello se añade que la progresiva dependencia tecnológica de las Administraciones Públicas puede favorecer fenómenos de automatización acrítica en la adopción de decisiones administrativas. La aparente objetividad técnica de los sistemas algorítmicos puede generar una confianza excesiva en los resultados automatizados, debilitando la capacidad de cuestionamiento y supervisión por parte de los operadores humanos.

Desde esta perspectiva, la supervisión humana debe integrarse dentro de modelos más amplios de *accountability* y gobernanza continua de los sistemas automatizados. El concepto de *accountability* no se limita únicamente a la posibilidad de exigir responsabilidad ex post, sino que implica también la existencia de mecanismos preventivos orientados a garantizar transparencia, trazabilidad y control permanente sobre las decisiones adoptadas mediante inteligencia artificial.

En este sentido, la *accountability* exige no solo la posibilidad de identificar responsables jurídicos, sino también la capacidad de explicar, justificar y revisar las decisiones automatizadas adoptadas por la Administración Pública. La legitimidad de los sistemas algorítmicos depende, en gran medida, de que los ciudadanos puedan comprender las razones que fundamentan una determinada decisión y disponer de mecanismos efectivos para cuestionarla o impugnarla. Como señala Wieringa, la accountability implica «the obligation to explain and justify decisions, as well as the capacity to contest and review them»[18].

En el ámbito europeo, el Reglamento de Inteligencia Artificial refuerza precisamente esta lógica preventiva mediante la incorporación de obligaciones

18 Wieringa, M., «What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability», *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, p. 7. Traducción propia: «La rendición de cuentas implica la obligación de explicar y justificar las decisiones, así como la capacidad de impugnarlas y someterlas a revisión».

vinculadas a supervisión humana, gestión de riesgos, documentación técnica y control continuo de determinados sistemas considerados de alto riesgo. Este enfoque refleja una evolución progresiva desde modelos tradicionales de control administrativo hacia esquemas de gobernanza tecnológica más dinámicos y adaptativos.

Asimismo, la consolidación de modelos efectivos de *accountability* exige desarrollar estructuras institucionales capaces de supervisar de manera continuada el funcionamiento de los sistemas automatizados utilizados por las Administraciones Públicas. La complejidad técnica y el impacto potencial de estas tecnologías hacen necesario avanzar hacia mecanismos especializados de evaluación, auditoría y control interdisciplinar que integren perspectivas jurídicas, técnicas y organizativas.

En consecuencia, puede afirmarse que la supervisión humana y la *accountability* constituyen elementos estructurales de la gobernanza algorítmica en el ámbito público. La utilización legítima de sistemas de inteligencia artificial en la Administración Pública depende no solo de la existencia de marcos normativos adecuados, sino también de la capacidad efectiva de las instituciones para garantizar control, transparencia y responsabilidad sobre las decisiones automatizadas.

V. Propuesta de gobernanza algorítmica

La superación de los desafíos identificados en la implementación de la inteligencia artificial en la Administración Pública exige la construcción de un modelo de gobernanza algorítmica que integre de manera coherente dimensiones jurídicas, técnicas y organizativas. No se trata únicamente de reforzar el marco normativo existente, sino de articular mecanismos que permitan su aplicación efectiva en contextos reales de decisión automatizada.

En este sentido, la gobernanza de los sistemas de inteligencia artificial debe configurarse como un proceso continuo, orientado no solo al cumplimiento normativo, sino también a la gestión dinámica de riesgos. Como ha señalado la doctrina, la regulación de estos sistemas no puede limitarse a un enfoque estático, sino que requiere estructuras de control adaptativas que permitan responder a la evolución constante de la tecnología[19].

Desde esta perspectiva, uno de los pilares fundamentales del modelo propuesto es la integración del *compliance* digital como mecanismo estructural de control. La adopción de sistemas de inteligencia artificial en el ámbito público debe ir acompañada de procedimientos internos destinados a garantizar el cumplimiento de los principios de legalidad, transparencia y responsabilidad. En este contexto, se ha señalado que las organizaciones que implementan sistemas algorítmicos deben adoptar marcos de gobernanza que garanticen mecanismos de supervisión continua y rendición de cuentas[20].

19 Veale, M., Brass, I., «Administration by algorithm? Public management meets public sector machine learning», *Public Management Review*, 2019.

20 Rodríguez, P., *Algorithmic governance in the public sector: risks and challenges*, 2020.

Asimismo, resulta imprescindible reforzar los mecanismos de auditoría algorítmica, no solo como herramientas de control ex post, sino como instrumentos de evaluación previa y continua del funcionamiento de los sistemas. Estas auditorías deben permitir identificar riesgos potenciales antes de la implementación y verificar el comportamiento del sistema durante su funcionamiento. En este sentido, la literatura reciente ha destacado que «algorithmic systems require continuous auditing to detect and mitigate unintended consequences»[21].

Otro elemento esencial del modelo de gobernanza propuesto es la consolidación de una supervisión humana efectiva, entendida no como una intervención meramente formal, sino como una capacidad real de control sobre las decisiones automatizadas. La supervisión humana debe garantizar la posibilidad de intervenir, corregir o incluso anular decisiones cuando estas resulten incompatibles con el ordenamiento jurídico. En este sentido, se ha señalado que «human oversight mechanisms must be designed to enable meaningful intervention in automated decision-making processes»[22].

Desde una perspectiva institucional, la gobernanza algorítmica exige también la adaptación de las estructuras administrativas. La incorporación de inteligencia artificial en la Administración Pública requiere la creación de unidades especializadas, la formación del personal y el desarrollo de capacidades técnicas que permitan comprender y gestionar estos sistemas. Como ha sido señalado, la implementación efectiva de la inteligencia artificial depende en gran medida de la capacidad institucional para integrarla de forma adecuada en los procesos administrativos[23].

Finalmente, la construcción de un modelo de gobernanza algorítmica debe incorporar una dimensión preventiva, orientada a anticipar los riesgos asociados al uso de estas tecnologías. La identificación temprana de posibles impactos en derechos fundamentales, así como la adopción de medidas correctivas antes de la implementación, resulta esencial para garantizar una utilización responsable de la inteligencia artificial en el ámbito público.

A partir de lo anterior, puede sostenerse que la gobernanza algorítmica no debe concebirse como un elemento accesorio, sino como un componente estructural del proceso de digitalización de la Administración Pública. Solo a través de la integración efectiva de mecanismos jurídicos, técnicos y organizativos será posible garantizar que el uso de la inteligencia artificial se alinee con los principios del Estado de Derecho y con la protección de los derechos fundamentales.

21 Dieterich, W., Mendoza, C., Brenner, M., *COMPAS Risk Scales: Demonstrating Accuracy Equity*, 2016. Traducción propia: «Los sistemas algorítmicos requieren auditorías continuas para detectar y mitigar consecuencias no intencionadas».

22 High-Level Expert Group on AI, *Policy and Investment Recommendations for Trustworthy AI*, 2019. Traducción propia: «Los mecanismos de supervisión humana deben diseñarse de modo que permitan una intervención significativa en los procesos de toma de decisiones automatizadas».

23 OECD, *Artificial Intelligence in Society*, 2019.

VI. Conclusiones

El análisis desarrollado permite afirmar que la incorporación de sistemas de inteligencia artificial en la Administración Pública no constituye un fenómeno meramente tecnológico, sino una transformación estructural del ejercicio del poder público que exige una reformulación de las categorías jurídicas tradicionales. La automatización de decisiones con impacto en derechos fundamentales introduce una nueva dimensión en la actuación administrativa, en la que la opacidad algorítmica y la complejidad técnica tensionan los principios clásicos de transparencia, motivación y control.

El modelo normativo europeo, articulado en torno al RIA y su interacción con el RGPD y la Directiva NIS2, representa un avance significativo en la construcción de un marco jurídico orientado a la gestión de riesgos. No obstante, se evidencia la persistencia de una brecha entre regulación e implementación que limita la eficacia real de los mecanismos de control, poniendo de manifiesto que la densificación normativa no garantiza por sí misma una aplicación efectiva.

La utilización de sistemas algorítmicos en la toma de decisiones administrativas genera tensiones relevantes en materia de responsabilidad jurídica, al diluir la imputación en contextos caracterizados por la participación de múltiples actores y por la dificultad de reconstrucción del proceso decisional. Esta realidad obliga a reinterpretar los fundamentos del Derecho administrativo, manteniendo la centralidad de la responsabilidad pública, pero adaptando sus mecanismos a la lógica distribuida y dinámica de la inteligencia artificial.

La efectividad de la gobernanza algorítmica depende de la articulación de mecanismos operativos de control que superen el cumplimiento formal de las normas. En este sentido, la auditoría algorítmica, la supervisión humana significativa y la integración del *compliance* digital se configuran como elementos esenciales para garantizar la legalidad, la legitimidad y la confianza en el uso de estas tecnologías en el ámbito público.

La progresiva sofisticación de los sistemas de inteligencia artificial utilizados en el ámbito público exige avanzar hacia modelos permanentes de supervisión preventiva y evaluación continua, capaces de garantizar no solo el cumplimiento normativo inicial, sino también el control dinámico del comportamiento algorítmico durante todo el ciclo de vida del sistema. La gobernanza algorítmica requiere, por tanto, estructuras institucionales estables de auditoría, trazabilidad y supervisión interdisciplinar que permitan responder adecuadamente a la evolución constante de estas tecnologías.

La gobernanza de la inteligencia artificial en la Administración Pública debe orientarse hacia un modelo preventivo, dinámico e integrado, capaz de anticipar riesgos y adaptarse a la evolución tecnológica. La construcción de este modelo no constituye únicamente una exigencia jurídica, sino una condición necesaria para preservar los principios del Estado de Derecho, la legitimidad democrática y la protección efectiva de los derechos fundamentales en un contexto de creciente automatización del poder público.

En consecuencia, el verdadero desafío de la inteligencia artificial en la Administración Pública no reside únicamente en la capacidad de automatizar decisiones, sino en garantizar que la transformación tecnológica del poder público continúe sometida a los principios democráticos, al control jurídico y a las garantías propias de un Estado social y democrático de Derecho.

Referencias bibliográficas

Ananny, M., Crawford, K. «Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability», *New Media & Society*, vol. 20, n.º 3 (2018), p. 979. https://www.researchgate.net/publication/312378887_Seeing_without_knowing_Limitations_of_the_transparency_ideal_and_its_application_to_algorithmic_accountability

Bovens, M. «Analysing and Assessing Accountability: A Conceptual Framework», *European Law Journal*, 2007. https://dspace.library.uu.nl/handle/1874/35005

Burr, C. y Cramer, H. The EU's Artificial Intelligence Act: Risks and Regulatory Challenges, 2018. https://arxiv.org/abs/1811.08856

Cerrillo i Martínez, A. «El impacto de la inteligencia artificial en el Derecho administrativo ¿nuevos conceptos para nuevos retos tecnológicos?», *Revista General de Derecho Administrativo*, n.º 50 (2019). https://dialnet.unirioja.es/servlet/articulo?codigo=6823517

Cerrillo i Martínez, A. «Retos y oportunidades del uso de la inteligencia artificial en las administraciones públicas», *RIGL: Revista Iberoamericana de Gobierno Local*, n.º 19 (2021). https://dialnet.unirioja.es/servlet/articulo?codigo=9412440

Cerrillo i Martínez, A. «Transparencia administrativa y uso de algoritmos por las Administraciones públicas», *Revista Catalana de Dret Públic*, n.º 58 (2019). https://dialnet.unirioja.es/servlet/articulo?codigo=8021169

Comisión Europea. *Ethics Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence*, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Dieterich, W., Mendoza, C., Brenner, M. *COMPAS Risk Scales: Demonstrating Accuracy Equity*, 2016. https://arxiv.org/abs/1603.07707

Gamero Casado, E. «Necesidad de motivación e invalidez de los actos administrativos sustentados en inteligencia artificial o en algoritmos», Almacén de Derecho, 2021. Disponible en: https://almacendederecho.org/necesidad-de-motivacion-e-invalidez-de-los-actos-administrativos-sustentados-en-inteligencia-artificial-o-en-algoritmos

High-Level Expert Group on AI. *Policy and Investment Recommendations for Trustworthy AI*, 2019. https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., Yu, H., «Accountable Algorithms», *University of Pennsylvania Law Review*, vol. 165 (2017), 633-705. https://repository.law.upenn.edu/documents?sear-

ch=%22https%3A%2F%2Fscholarship.law.upenn.edu%2Fpenn_law_review%2Fvol165%2Fiss3%2F3%2F%22&searchtypes=Metadata|Full%20text&applyState=true

Kuner, C., Bygrave, L. A., Docksey, C. (Eds.). *The EU General Data Protection Regulation (GDPR): A Commentary — 2021 Update*. Maastricht Faculty of Law Working Paper, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3839645

OECD *et al. Latin American Economic Outlook 2023: Investing in Sustainable Development*. OECD Publishing, 2023. https://doi.org/10.1787/8c93ff6e-en

OECD. *Artificial Intelligence in Society*, 2019. https://www.oecd.org/en/topics/digital.html

Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015. https://raley.english.ucsb.edu/wp-content/Engl800/Pasquale-blackbox.pdf

Piñar Mañas, J. L. «Revolución tecnológica, Derecho administrativo y Administración Pública: notas provisionales para una reflexión», *DIXI*, n.º 13 (2011), 8-32.

Sunstein, C. R. *Algorithms, Correcting Biases, and Public Policy*. Harvard Law School, 2018. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2992473

Unión Europea. *Directiva (UE) 2022/2555 del Parlamento Europeo y del Consejo de 14 de diciembre de 2022 relativa a las medidas destinadas a garantizar un elevado nivel común de ciberseguridad en toda la Unión, por la que se modifican el Reglamento (UE) n.º 910/2014 y la Directiva (UE) 2018/1972 y por la que se deroga la Directiva (UE) 2016/1148 (Directiva SRI 2)*. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32022L2555

Unión Europea. *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)*. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32016R0679

Unión Europea. *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*. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689

Veale, M., Borgesius, F. Z. «Demystifying the Draft EU Artificial Intelligence Act», *Computer Law Review International*, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3896852

Veale, M., Brass, I. «Administration by algorithm? Public management meets public sector machine learning», *Public Management Review*, 2019. https://arxiv.org/abs/1807.04206

Wachter, S., Mittelstadt, B., Floridi, L. «Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation»,

International Data Privacy Law, vol. 7, n.º 2 (2017): 76-99. https://ssrn.com/abstract=2903469

Wieringa, M. «What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability», *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, p. 7. https://dl.acm.org/doi/10.1145/3351095.3372833

Yeung, K. «Algorithmic regulation: A critical interrogation», *Regulation & Governance*, vol. 12, n.º 4 (2018), 505-523. https://onlinelibrary.wiley.com/doi/10.1111/rego.12158

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 46-59

IMPACTO DE LA INTELIGENCIA ARTIFICIAL EN LA PROTECCIÓN JURÍDICA DE LOS DERECHOS FUNDAMENTALES

IMPACT OF ARTIFICIAL INTELLIGENCE ON THE LEGAL PROTECTION OF FUNDAMENTAL RIGHTS

M.ª Olivia Galán Azofra

Magistrada suplente, Audiencia Provincial Salamanca

Resumen

El artículo analiza el impacto de la inteligencia artificial en la protección de los derechos fundamentales, centrándose especialmente en la igualdad, la privacidad y el debido proceso. Se examinan riesgos como la opacidad algorítmica, la reproducción de sesgos presentes en los datos de entrenamiento y la automatización de decisiones sin supervisión humana suficiente, factores que pueden generar discriminación, limitar la protección de datos personales y dificultar el ejercicio del derecho de defensa. Además, se comparan los principales marcos regulatorios, especialmente el AI Act de la Unión Europea, el Convenio Marco del Consejo de Europa y la integración entre el GDPR y las evaluaciones de impacto en derechos fundamentales. Se afirma que los mecanismos normativos actuales representan avances significativos, pero todavía resultan insuficientes para afrontar los desafíos derivados de sistemas cada vez más complejos, por lo que se requiere reforzar la transparencia, la supervisión humana y la rendición de cuentas.

Palabras clave

Inteligencia artificial, derechos fundamentales, discriminación algorítmica, privacidad y protección de datos, transparencia y supervisión jurídica

Abstract

The article examines the impact of artificial intelligence on the protection of fundamental rights, with particular emphasis on equality, privacy, and due process. It analyzes risks such as algorithmic opacity, the reproduction of biases embedded in training data, and the automation of decision-making processes without sufficient human oversight. These factors may lead to discrimination, undermine the protection of personal data, and hinder the effective exercise of the right to defense. Furthermore, the article compares the main regulatory frameworks, particularly the European Union's AI Act, the Council of Europe Framework Convention on Artificial Intelligence, and the integration of the GDPR with Fundamental Rights Impact Assessments. It argues that current regulatory mechanisms represent significant progress; however, they remain insufficient to address the challenges posed by increasingly complex AI systems. Consequently, there is a need to strengthen transparency, human oversight, and accountability mechanisms.

Keywords

Artificial intelligence, fundamental rights, algorithmic discrimination, privacy and data protection, transparency and accountability

I. Introducción

En los últimos años, la inteligencia artificial ha empezado a ocupar un espacio que antes parecía reservado exclusivamente a la decisión humana. No siempre de forma visible, pero sí constante. Desde procesos administrativos hasta ámbitos más sensibles como el sistema penal o el acceso al empleo, los algoritmos han ido asumiendo funciones que inciden, de manera directa o indirecta, en la vida de las personas. Esto, que en principio podría interpretarse como un avance en eficiencia, plantea dudas más profundas cuando se analiza desde el prisma jurídico. Este fenómeno plantea implicaciones directas en la delimitación de la responsabilidad jurídica y en la eficacia de las garantías constitucionales.

El problema no radica tanto en la tecnología en sí, sino en cómo se integra en estructuras que, históricamente, han sido diseñadas para garantizar derechos. En la práctica, muchos sistemas de inteligencia artificial operan a partir de datos previos que no son neutrales. Arrastran sesgos, desigualdades y patrones difíciles de detectar a simple vista. De ahí que, como apunta Domínguez Álvarez[1], las decisiones automatizadas no puedan entenderse únicamente como procesos técnicos, sino como actuaciones con efectos jurídicos relevantes, especialmente en relación con el derecho a la protección de datos, especialmente a la luz de las exigencias de explicabilidad recogidas en los desarrollos normativos europeos sobre inteligencia artificial.

A esto se suma otra cuestión que complica aún más el escenario: la falta de transparencia. No siempre es posible conocer cómo un algoritmo llega a una determinada conclusión. Y cuando esa conclusión afecta a derechos, por ejemplo, al acceso a un servicio o a la valoración de un riesgo, la imposibilidad de comprender el proceso debilita la capacidad de defensa. En este sentido, Miró Llinares[2] advierte que el verdadero riesgo no está solo en el posible error del sistema, sino en la dificultad para cuestionarlo, precisamente por esa apariencia de objetividad que muchas veces se le atribuye. La discriminación algorítmica no responde a una voluntad subjetiva, sino a la reproducción estructural de patrones presentes en los datos de entrenamiento.

Si se lleva esta reflexión un paso más allá, el impacto sobre el principio de igualdad también resulta evidente. La inteligencia artificial no discrimina de forma intencionada, pero puede hacerlo de manera estructural. Esto genera una forma de desigualdad más sutil, menos visible, pero igualmente problemática. Sánchez Díaz[3] subraya que estas nuevas formas de discriminación no encajan fácilmente en las categorías jurídicas tradicionales, lo que obliga a replantear los mecanismos de protección existentes.

1 Domínguez Álvarez, J. L. «Inteligencia Artificial, derecho administrativo y protección de datos personales. Entre la dignidad de la persona y la eficacia administrativa», *Ius et Scientia,* vol. 1, n.º 7 (2021), pp. 304-326.

2 Miró Llinares, F. «Inteligencia artificial y Justicia Penal. Más allá de los resultados lesivos causados por robots», *Revista de Derecho Penal y Criminología*, n.º 20 (2018), pp. 87-130.

3 Sánchez Díaz, M. F. «El impacto de la inteligencia artificial generativa en los derechos humanos», *Revista Eurolatinoamericana de Derecho Administrativo,* vol. 11, n.º 1 (2024).

Con este panorama, surge una inquietud que atraviesa todo el análisis: no está del todo claro hasta qué punto los marcos normativos actuales son capaces de responder a estos desafíos. La inteligencia artificial avanza con rapidez, mientras que el Derecho, por su propia naturaleza, se mueve con mayor cautela. Esa diferencia de ritmo abre un espacio de incertidumbre que merece ser examinado con detenimiento.

Es precisamente en ese punto donde se sitúa la pregunta de investigación que guía este trabajo: *¿De qué manera la implementación de sistemas de inteligencia artificial incide en la protección y garantía de los derechos fundamentales, especialmente en relación con la igualdad, la privacidad y el debido proceso?*

Partiendo de esta cuestión, se plantea que *la ausencia de marcos regulatorios sólidos y mecanismos de supervisión efectivos en el uso de inteligencia artificial puede comprometer la garantía de los derechos fundamentales, especialmente la igualdad, la privacidad y el debido proceso.* En particular, la opacidad de los algoritmos, la reproducción de sesgos y la automatización de decisiones sin intervención humana suficiente tienden a debilitar garantías como la igualdad, la privacidad y el debido proceso.

A partir de esta hipótesis, el trabajo se orienta a analizar la incidencia de la inteligencia artificial en los derechos fundamentales mediante un estudio doctrinal y normativo del marco jurídico vigente, prestando especial atención a los riesgos asociados a la toma de decisiones automatizadas, identificando criterios que permitan reforzar la protección de la dignidad humana y las garantías propias del Estado de Derecho.

Para alcanzar este propósito general, el análisis se articula en tres líneas complementarias. En primer lugar, se busca definir los conceptos de inteligencia artificial y derechos fundamentales, delimitando su relación en el contexto jurídico actual, a partir de la revisión de la doctrina y la normativa existente. En segundo lugar, se pretende analizar el impacto concreto de los sistemas de inteligencia artificial sobre derechos como la igualdad, la privacidad y el debido proceso, atendiendo no solo a sus implicaciones teóricas, sino también a sus efectos prácticos, sus riesgos y sus limitaciones técnicas. Finalmente, se propone comparar los distintos enfoques regulatorios en materia de inteligencia artificial, identificando puntos de coincidencia y divergencia en la forma en que abordan la protección de los derechos fundamentales, con el objetivo de aportar criterios que puedan servir de base para futuras mejoras normativas.

II. Inteligencia artificial y derechos fundamentales: delimitación conceptual y aproximación doctrinal

Antes de analizar impactos o riesgos, conviene detenerse en algo más básico, aunque no siempre sencillo: qué entendemos por inteligencia artificial y por derechos fundamentales. No es una cuestión meramente terminológica. En realidad, de cómo se definan ambos conceptos depende, en buena medida, la forma en que se interpretan sus puntos de contacto. La doctrina jurídica ha abordado estas nociones desde perspectivas diversas, y eso explica que no exista una única

definición cerrada, sino aproximaciones que responden a contextos y preocupaciones distintas.

1. Concepto de inteligencia artificial desde la doctrina

Por su parte, Navas Navarro[4] ofrece una definición más próxima al Derecho civil, entendiendo la inteligencia artificial como sistemas tecnológicos que simulan procesos cognitivos humanos y que pueden actuar de forma autónoma en determinados contextos. Su aportación resulta interesante porque introduce la idea de simulación de capacidades humanas, lo que acerca el debate a cuestiones clásicas del Derecho, como la imputación de conductas. En la práctica, esto abre preguntas incómodas: si un sistema actúa «como si pensara», ¿hasta qué punto puede ser tratado como un mero instrumento? Esta aproximación permite vincular la IA con categorías clásicas del Derecho civil, aunque resulta limitada al no incorporar plenamente su dimensión sistémica en la producción de decisiones automatizadas.

Desde una perspectiva más centrada en la gobernanza digital, Hildebrandt propone entender la inteligencia artificial como sistemas computacionales que generan inferencias a partir de datos, influyendo en decisiones que afectan a personas[5]. Su enfoque introduce un matiz relevante: la IA no solo decide, sino que condiciona entornos. Esto sugiere que su impacto no siempre es directo, pero sí estructural. En el ámbito jurídico, esta idea resulta especialmente útil porque permite entender cómo la IA puede influir en derechos fundamentales sin necesidad de una intervención explícita o visible. Esta concepción resulta especialmente relevante al desplazar el foco desde la decisión individual hacia los efectos estructurales de la tecnología.

Otra aproximación bastante clara es la que ofrece Pérez-Ugena[6], quien señala que la inteligencia artificial puede entenderse como un sistema basado en máquinas capaz de procesar información y generar resultados, como predicciones, recomendaciones o decisiones, que influyen en entornos reales o virtuales. Lo interesante de su planteamiento es que recoge, en cierta medida, la definición que se está consolidando en el ámbito europeo, especialmente a partir de la regulación reciente. Esto introduce un matiz importante: la inteligencia artificial no se define solo por su funcionamiento técnico, sino por su capacidad de producir efectos. En la práctica, esto implica que el Derecho no puede limitarse a regular la herramienta, sino que debe atender a las consecuencias jurídicas de su uso, que pueden afectar directamente a derechos fundamentales.

Las distintas aproximaciones coinciden en la centralidad de la dignidad humana, aunque difieren en el grado de normativización y exigibilidad jurídica de los derechos fundamentales.

4 Navas Navarro, S. «Responsabilidad civil e Inteligencia artificial», *El Cronista del Estado Social y Democrático de Derecho*, n.º 100 (2022), pp. 106-115.

5 Hildebrandt, M. «The Artificial Intelligence of European Union Law», *German Law Journal,* vol. 21, n.º 1 (2020), pp. 74-79.

6 Pérez-Ugena, M. «La inteligencia artificial: definición, regulación y riesgos para los derechos fundamentales», *Estudios de Deusto,* vol. 72, n.º 1 (2024), pp. 307-337.

2. Concepto de derechos fundamentales desde la doctrina

Una de las aproximaciones más claras y rigurosas proviene de Ibáñez Macías[7], quien en su artículo sobre la identificación de derechos fundamentales en la Constitución española explica que estos no son meras categorías dogmáticas, sino posiciones jurídicas que vinculan a todos los poderes públicos y se extienden más allá de un listado literal. Es decir, el concepto de derechos fundamentales tiene que ver con la fuerza normativa que poseen dentro del ordenamiento constitucional: su carácter no es opcional ni decorativo, sino que obliga a los poderes públicos a respetarlos y garantizarlos. Ibáñez Macías muestra cómo esta fundamentalidad obliga a que las normas que los reconocen no solo existan en la letra, sino que se traduzcan en conductas obligatorias para el Estado y sus instituciones.

Otra manera de aproximarse al concepto se encuentra en el trabajo de Gonzales Ojeda[8], quien describe los derechos fundamentales como aquellos derechos subjetivos públicos que tienen una doble dimensión: por un lado, su reconocimiento como libertades esenciales de la persona, y por otro, su protección mediante mecanismos constitucionales especiales. Esta perspectiva combina dos elementos: la dimensión moral de los derechos con su eficacia jurídica real. Gonzales Ojeda incluso contrasta distintas expresiones conceptuales, derechos humanos, libertades públicas, derechos naturales, para concluir que la idea de «derechos fundamentales» es la más adecuada cuando se quiere enfatizar su fuerza jurídica positiva dentro del sistema constitucional.

Desde una óptica doctrinal más general, el artículo de Nogueira Alcalá[9] ofrece una definición detallada al afirmar que los derechos fundamentales son atributos esenciales de la dignidad humana, consagrados en las constituciones modernas y dotados de mecanismos especiales de garantía. Para Nogueira Alcalá, su relevancia va más allá de una lista de libertades o potestades: constituyen límites al poder estatal y están directamente vinculados con el contenido esencial de la persona como sujeto de derecho. Esto quiere decir que su protección no depende de la voluntad de quien detenta el poder, sino de su reconocimiento como núcleo inalienable dentro del orden jurídico.

3. Interacción en el contexto jurídico actual

En la práctica, hablar de inteligencia artificial y derechos fundamentales ya no es solo un debate teórico. Se ha convertido en un reto concreto para el Derecho constitucional y administrativo, justamente porque la IA se infiltra en

7 Ibáñez Macías, A. «Identificando derechos fundamentales en la Constitución española», *Derechos y libertades: Revista de Filosofía del Derecho y derechos humanos*, n.º 44 (2021), pp. 277-315.

8 Gonzales Ojeda, M. «Introducción a los derechos fundamentales y su protección constitucional», *Ius Inkarri: Revista de la Facultad de Derecho y Ciencia Política de la Universidad Ricardo Palma,* vol. 1, n.º 1 (2011), pp. 41-73.

9 Nogueira Alcalá, H. «Aspectos de una Teoría de los Derechos Fundamentales: La Delimitación, Regulación, Garantías y Limitaciones de los Derechos Fundamentales», *Revista Ius et Praxis,* vol. 11, n.º 2 (2005), pp. 15-64.

decisiones que afectan a la dignidad, la igualdad o la privacidad de las personas. En este punto, varios autores han analizado cómo la doctrina y los marcos normativos actuales intentan responder a estas tensiones.

Un primer abordaje interesante lo ofrecen Castellanos Claramunt y Montero Caro[10], quienes examinan la incidencia de sistemas algorítmicos en las resoluciones judiciales y cómo estos pueden influir sobre el derecho a la tutela judicial efectiva. Señalan que, si bien los sistemas basados en inteligencia artificial pueden agilizar procesos, también pueden introducir opacidad en la fundamentación de las decisiones, lo que complica el ejercicio de las garantías constitucionales tradicionales. En su análisis se considera la necesidad de que el Derecho adopte herramientas de control que permitan garantizar que la automatización no erosione las bases del debido proceso ni del acceso a la justicia.

Otra perspectiva relevante es la que plantea Aponte Fonseca[11], quien analiza las tensiones y realidades de la IA frente a los derechos fundamentales desde la falta de una regulación específica. En su estudio sobre el caso colombiano, muestra que la ausencia de normas claras facilita escenarios de vulneración de derechos, como la discriminación automatizada o la falta de transparencia en decisiones públicas o privadas. Su conclusión sugiere que el Derecho debe evolucionar de manera más proactiva, no solo reaccionar a las consecuencias negativas, sino anticipar mecanismos de protección eficaces.

Finalmente, en un enfoque comparado y más amplio, Bermúdez-Tapia y Cárdenas-Gonzales[12] exploran los desafíos y perspectivas de los derechos fundamentales en el contexto de la IA, enfatizando la necesidad de equilibrar la innovación tecnológica con la protección de las libertades básicas. Estos autores destacan que, más allá de la normativa interna de cada país, existe una dimensión global que obliga a países y sistemas jurídicos a coordinarse para asegurar estándares comunes, especialmente en cuanto a la privacidad, el sesgo algorítmico y la equidad. La discusión incluye la relevancia de herramientas como evaluaciones de impacto sobre derechos fundamentales y la participación pública en el desarrollo de políticas tecnológicas.

En este contexto, la principal dificultad no reside únicamente en la existencia de riesgos tecnológicos, sino en la capacidad del Derecho para traducir dichos riesgos en categorías jurídicas operativas.

10 Castellanos Claramunt, J. y Montero Caro, M. D. «Perspectiva constitucional de las garantías de aplicación de la Inteligencia Artificial: la ineludible protección de los derechos fundamentales», *Ius et Scientia,* vol. 6, n.º 2 (2020), pp. 72-82.

11 Aponte Fonseca, Y. «Tensiones y realidades sobre la vulneración de los derechos fundamentales a falta de regulación de la inteligencia artificial (IA) en Colombia», *Revista Doctrina Distrital*, vol.4, n.º 1 (2024), pp. 45-79.

12 Bermúdez-Tapia, M. y Cardenas-Gonzales, J. R. «Derechos Fundamentales e Inteligencia Artificial: Desafíos y Perspectivas», *Población y Desarrollo*, vol. 31, n.º 61 (2005).

III. Impacto de la inteligencia artificial en los derechos fundamentales

La introducción de sistemas de inteligencia artificial en decisiones que afectan la vida de las personas no solo plantea oportunidades, sino también tensiones con derechos fundamentales básicos. Estas tecnologías magnifican volúmenes de datos, automatizan decisiones y operan muchas veces en condiciones no transparentes, lo que puede tener efectos relevantes sobre la igualdad, la privacidad y el debido proceso. Entender estos impactos exige distinguir las formas en que la IA puede reproducir sesgos existentes, erosionar la autonomía individual o dificultar la protección efectiva de garantías constitucionales en contextos contemporáneos.

1. Igualdad y no discriminación: sesgos algorítmicos y su impacto

La igualdad ante la ley está en riesgo cuando los sistemas de inteligencia artificial reproducen o amplifican sesgos presentes en los datos. En un estudio reciente sobre derechos humanos y sesgos de género, Estrada Tanck[13] analiza cómo la IA tiende a reflejar desigualdades estructurales en contextos como el empleo, la educación o los servicios públicos. Los algoritmos, al aprender de bases de datos históricas, pueden reforzar estereotipos y producir discriminaciones indirectas que afectan de manera desproporcionada a grupos marginados, como mujeres y niñas. En la práctica, esto significa que, sin medidas de corrección técnica y jurídica, la IA puede perpetuar injusticias que el derecho pretende combatir. Este fenómeno obliga a replantear los estándares probatorios en materia de discriminación indirecta en entornos automatizados.

2. Privacidad y protección de datos: retos frente a la recopilación masiva

La privacidad se ve particularmente tensionada por la capacidad de la IA para procesar enormes cantidades de información personal de manera automática. En el análisis que realizan Martabit Sagredo y Peña[14], se evidencia que los sistemas de inteligencia artificial pueden vulnerar el derecho a la privacidad porque agrupan y cruzan datos sin un control eficaz, posibilitando prácticas discriminatorias o intrusivas. Su trabajo se basa en experiencias prácticas y en la comparación con marcos normativos como el europeo, destacando que la automatización pone en jaque la protección de datos si no existen salvaguardas claras. Esto sugiere que la IA puede entrar en conflicto con derechos fundamentales básicos si se deja sin regulación o sin criterios de proporcionalidad, lo que refuerza dinámicas de asimetría informativa entre administraciones, empresas tecnológicas y ciudadanos.

13 Estrada Tanck, D. «Inteligencia artificial y derechos humanos de las mujeres y las niñas: sesgos, desafíos y oportunidades a la luz del Derecho Internacional», *Cuadernos de Derecho Transnacional,* vol. 18, n.º 1 (2026), pp. 548-569.

14 Martabit Sagredo, M. J. y Peña, M. «Inteligencia artificial y derechos fundamentales: impacto en los derechos de la privacidad», *Revista de Derecho UDD,* vol. XXV, n.º 50 (2024).

3. Debido proceso y justicia: limitaciones técnicas y opacidad

El debido proceso exige que las personas afectadas por una decisión puedan comprender, cuestionar y, en caso necesario, impugnar la forma en que esa decisión fue adoptada. En su artículo sobre las garantías procesales en el contexto penal, Furingo[15] explora cómo la aplicación de sistemas basados en IA puede amenazar estos principios clásicos. Señala que la falta de explicabilidad de muchos algoritmos, la dificultad para ofrecer motivos comprensibles y la ausencia de espacios de participación humana efectiva pueden erosionar derechos como la contradictoriedad o la tutela judicial efectiva. En suma, cuando las decisiones automatizadas reemplazan o reducen la intervención humana significativa sin mecanismos de control adecuados, se pone en riesgo la esencia misma del debido proceso. En consecuencia, la opacidad algorítmica puede tensionar el núcleo esencial del derecho de defensa en entornos digitalizados.

IV. Comparación de enfoques regulatorios en materia de inteligencia artificial y derechos fundamentales

El auge de la regulación de la inteligencia artificial ha generado distintos modelos jurídicos que intentan equilibrar innovación y protección de derechos fundamentales. Mientras algunos marcos buscan establecer obligaciones claras con impacto directo en garantías básicas, otros operan de forma más amplia, incorporando principios generales o instrumentos evaluativos. A continuación, se comparan tres enfoques regulatorios predominantes, el *AI Act* de la Unión Europea, el enfoque del Consejo de Europa y el marco de protección de datos complementario, con énfasis en sus similitudes y diferencias en la salvaguarda de derechos fundamentales. La comparación se estructura en función de tres variables: nivel de obligatoriedad normativa, enfoque de protección de derechos y grado de concreción operativa.

Tabla 1. Comparación de enfoques regulatorios

Enfoque regulatorio	¿Qué es?	Protección de derechos fundamentales (cómo lo aborda)	Limitaciones / Críticas principales
AI Act (Unión Europea)	Marco legislativo aprobado en 2024 que regula el uso de sistemas de IA en función de niveles de riesgo[16].	Coloca los derechos fundamentales como eje del cowntrol, exige evaluaciones de impacto de derechos (*Fundamental Rights Impact Assessment*) y categorización de sistemas de alto riesgo que pueden afectar derechos como igualdad o privacidad.	Puede enfatizar en la lógica de «producto/seguridad» más que en un enfoque holístico de derechos, generando tensiones en su aplicación real.

15 Furingo, C. I. «La Inteligencia Artificial: ¿innovación o amenaza para las garantías procesales en la justicia penal?», *Revista de Derecho de la UNED*, n.º 36 (2025), pp. 101-132.

16 Gill-Pedro, E. «Fundamental Rights Impact Assessments in the EU's AI Act: A teleological and contextual analysis of the obligations of deployers», *EJLT,* vol. 16, n.º 3 (2025).

Enfoque regulatorio	¿Qué es?	Protección de derechos fundamentales (cómo lo aborda)	Limitaciones / Críticas principales
Convenio Marco del Consejo de Europa sobre IA	Instrumento internacional orientado a coordinar enfoques normativos entre Estados parte[17].	Busca establecer principios éticos y jurídicos comunes para garantizar derechos como la dignidad humana, no discriminación y protección de datos, promoviendo supervisión humana y responsabilidad.	Más general en su formulación; deja libertad a los Estados para implementar medidas concretas, lo que puede generar divergencias normativas.
Protección de datos y evaluación de impacto integrada (GDPR y FRIA conjunta)	Integración de evaluación de impacto de datos y derechos fundamentales en sistemas de IA[18].	Combina análisis de riesgos de tratamiento de datos y de derechos fundamentales, proponiendo un enfoque sistemático que puede mejorar la protección de la privacidad y otras libertades conectadas.	Su aplicación puede resultar compleja por la superposición de procedimientos y obligaciones técnicas que no siempre se traducen en protección efectiva en todos los contextos.

Fuente: elaboración propia (2026).

1. Enfoques regulatorios con autores reales

La regulación de la inteligencia artificial ha generado distintos enfoques que buscan equilibrar innovación tecnológica y protección de derechos fundamentales. Algunos marcos, como el *AI Act* europeo, priorizan la obligatoriedad y la categorización por riesgos; otros, como los lineamientos del Consejo de Europa, enfatizan principios éticos y coordinación internacional. A su vez, la integración con normas de protección de datos, como el GDPR, añade un nivel técnico que refuerza la salvaguarda de derechos como privacidad y debido proceso. Esta sección analiza estas corrientes mediante autores reales, destacando similitudes, diferencias y áreas de mejora jurídica.

AI Act y protección de derechos fundamentales. El *Artificial Intelligence Act* aprobado por la Unión Europea se ha convertido en el primer intento formal y exhaustivo de regulación legal de IA con impacto directo sobre derechos humanos. Palmiotto[19] explica que el proceso legislativo del *AI Act* no solo busca armonizar el mercado interno europeo, sino también integrar la protección de derechos fundamentales como parte de su estructura normativa, con una evolución influenciada por tensiones políticas y compromisos entre instituciones europeas. Este enfoque resalta cómo la inclusión de estos derechos en la regulación técnica

17 Díaz Galán, E. «La regulación jurídica de la inteligencia artificial (IA) en Europa: dos pasos decisivos, aunque insuficientes, del Consejo de Europa y la Unión Europea», *Cuadernos de Derecho Transnacional,* vol. 17, n.º 1 (2025), pp. 366-390.

18 Thomaidou, A. y Limniotis, K. «Navigating Through Human Rights in AI: Exploring the Interplay Between GDPR and Fundamental Rights Impact Assessment», *J. Cybersecur. Priv.,* vol. 5, n.º 1 (2025), p. 7.

19 Palmiotto, F. *The AI Act Roller Coaster: The Evolution of Fundamental Rights Protection in the Legislative Process and the Future of the Regulation*, Cambridge Core, 2025.

pretende ofrecer garantías más amplias, aunque su eficacia dependerá de la implementación práctica y la interpretación judicial futura.

Marcos internacionales: Consejo de Europa y UE. El trabajo de Díaz Galán[20] examina la regulación jurídica de la IA desde una perspectiva europea más amplia, destacando los avances tanto del Consejo de Europa como de la Unión Europea. Este análisis muestra que, si bien ambos instrumentos comparten la preocupación por los derechos fundamentales, difieren en su carácter: el reglamento europeo es vinculante y detallado, mientras que el marco del Consejo de Europa pone mayor énfasis en principios éticos y coordinación entre Estados. La comparación permite observar que la protección de derechos en el contexto de IA puede ser más efectiva con instrumentos legalmente exigibles, aunque la cooperación internacional sigue siendo clave para estándares globales.

Protección de datos y derechos fundamentales integrados. Thomaidou y Limniotis[21] analizan el cruce entre la regulación de datos (GDPR) y la evaluación de impacto de derechos fundamentales (FRIA) en sistemas de IA. Este enfoque comparado muestra que la protección de la privacidad no puede entenderse de manera aislada y debe armonizarse con obligaciones más amplias de evaluación de riesgos sobre derechos humanos. Según estos autores, la integración de ambas evaluaciones puede ofrecer un modelo más coherente de protección en contextos donde la IA procesa grandes cantidades de datos personales, aunque la complejidad de su aplicación técnica puede limitar su eficacia si no se acompaña de mecanismos claros de supervisión.

V. Conclusiones

La incorporación de sistemas de inteligencia artificial en la toma de decisiones jurídicas, administrativas y sociales supone una transformación estructural en la forma en que se ejercen y protegen los derechos fundamentales. El análisis realizado permite afirmar que no se trata de un fenómeno meramente técnico, sino de un cambio que incide directamente en la arquitectura de garantías propias del Estado de Derecho, especialmente en lo relativo a la igualdad, la privacidad y el debido proceso.

La progresiva automatización de decisiones públicas y privadas está trasladando capacidad de decisión desde sujetos jurídicamente responsables hacia sistemas técnicos difíciles de controlar. Esto supone una transformación relevante en la estructura clásica de imputación y responsabilidad jurídica, en la medida en que la toma de decisiones deja de estar plenamente anclada en la intervención humana directa.

20 Díaz Galán, E. «La regulación jurídica de la inteligencia artificial (IA) en Europa: dos pasos decisivos, aunque insuficientes, del Consejo de Europa y la Unión Europea», *Cuadernos de Derecho Transnacional*, vol. 17, n.º 1 (2025), pp. 366-390.

21 Thomaidou, A. y Limniotis, K. «Navigating Through Human Rights in AI: Exploring the Interplay Between GDPR and Fundamental Rights Impact Assessment», *J. Cybersecur. Priv.,* vol. 5, n.º 1 (2025), p. 7.

En este sentido, la igualdad ante la ley se ve comprometida cuando los sistemas de inteligencia artificial reproducen patrones discriminatorios presentes en los datos de entrenamiento. La aparente neutralidad algorítmica puede ocultar formas de discriminación indirecta que resultan difíciles de identificar y corregir mediante las categorías jurídicas tradicionales, lo que exige una revisión de los instrumentos de tutela antidiscriminatoria.

La principal amenaza no reside únicamente en el error algorítmico, sino en la imposibilidad práctica de fiscalizar procesos automatizados complejos. La opacidad de los sistemas de inteligencia artificial dificulta no solo la comprensión de las decisiones adoptadas, sino también su control jurídico efectivo, debilitando así las garantías propias del debido proceso y la posibilidad real de defensa.

Del mismo modo, la privacidad se ve intensamente afectada por la capacidad de la inteligencia artificial para tratar grandes volúmenes de datos personales de forma masiva, continua y, en muchos casos, imperceptible para el individuo. Esta dinámica incrementa los riesgos de vigilancia y perfilado, lo que exige reforzar los mecanismos de control y supervisión más allá de la mera previsión normativa.

Los marcos regulatorios actuales constituyen avances relevantes en la protección de los derechos fundamentales frente a la inteligencia artificial, especialmente en el ámbito europeo. Sin embargo, estos instrumentos resultan todavía insuficientes para garantizar una protección plenamente efectiva ante sistemas de alta complejidad técnica, lo que evidencia la necesidad de reforzar los mecanismos de supervisión, control y rendición de cuentas.

En términos generales, la relación entre inteligencia artificial y derechos fundamentales se caracteriza por una tensión estructural entre innovación tecnológica y garantías jurídicas. El principal desafío no consiste únicamente en regular la tecnología, sino en preservar la centralidad de la persona frente a sistemas de decisión cada vez más automatizados. Ello exige reforzar la transparencia algorítmica, garantizar una intervención humana significativa en decisiones relevantes y desarrollar marcos de responsabilidad claros que permitan responder jurídicamente ante posibles vulneraciones.

Finalmente, el desarrollo de la inteligencia artificial obliga a repensar las categorías jurídicas tradicionales, no para sustituirlas, sino para adaptarlas a un entorno en el que las decisiones ya no provienen exclusivamente de sujetos humanos identificables. La protección efectiva de los derechos fundamentales dependerá, en última instancia, de la capacidad del Derecho para mantener su función de límite frente al poder, incluso cuando este se manifieste a través de sistemas técnicos complejos y automatizados.

Referencias bibliográficas

Aponte Fonseca, Y. «Tensiones y realidades sobre la vulneración de los derechos fundamentales a falta de regulación de la inteligencia artificial (IA) en Colombia», *Revista Doctrina Distrital*, vol. 4, n.º 1 (2024): 45-79.

Bermúdez-Tapia, M., Cardenas-Gonzales, J. R. «Derechos Fundamentales e Inteligencia Artificial: Desafíos y Perspectivas», *Población y Desarrollo*, vol. 31, n.º 61 (2005).

Castellanos Claramunt, J., Montero Caro, M. D. «Perspectiva constitucional de las garantías de aplicación de la Inteligencia Artificial: la ineludible protección de los derechos fundamentales», *Ius et Scientia*, vol. 6, n.º 2 (2020): 72-82.

Díaz Galán, E. «La regulación jurídica de la inteligencia artificial (IA) en Europa: dos pasos decisivos, aunque insuficientes, del Consejo de Europa y la Unión Europea», *Cuadernos de Derecho Transnacional*, vol. 17, n.º 1 (2025): 366-390.

Domínguez Álvarez, J. L. «Inteligencia Artificial, derecho administrativo y protección de datos personales. Entre la dignidad de la persona y la eficacia administrativa», *Ius et Scientia*, vol. 1, n.º 7 (2021): 304-326.

Estrada Tanck, D. «Inteligencia artificial y derechos humanos de las mujeres y las niñas: sesgos, desafíos y oportunidades a la luz del Derecho Internacional», *Cuadernos de Derecho Transnacional*, vol. 18, n.º 1 (2026): 548-569.

Furingo, C. I. «La Inteligencia Artificial: ¿innovación o amenaza para las garantías procesales en la justicia penal?», *Revista de Derecho de la UNED*, n.º 36 (2025): 101-132.

Gill-Pedro, E. «Fundamental Rights Impact Assessments in the EU's AI Act: A teleological and contextual analysis of the obligations of deployers», *EJLT*, vol. 16, n.º 3 (2025).

Gonzales Ojeda, M. «Introducción a los derechos fundamentales y su protección constitucional», *Ius Inkarri: Revista de la Facultad de Derecho y Ciencia Política de la Universidad Ricardo Palma*, vol. 1, n.º 1 (2011): 41-73.

Hildebrandt, M. «The Artificial Intelligence of European Union Law», *German Law Journal*, vol. 21, n.º 1 (2020): 74-79.

Ibáñez Macías, A. «Identificando derechos fundamentales en la Constitución española», *Derechos y libertades: Revista de Filosofía del Derecho y derechos humanos*, n.º 44 (2021): 277-315.

Martabit Sagredo, M. J., Peña, M. «Inteligencia artificial y derechos fundamentales: impacto en los derechos de la privacidad», *Revista de Derecho UDD*, vol. XXV, n.º 50 (2024).

Miró Llinares, F. «Inteligencia artificial y Justicia Penal. Más allá de los resultados lesivos causados por robots», *Revista de Derecho Penal y Criminología*, n.º 20 (2018): 87-130.

Navas Navarro, S. «Responsabilidad civil e Inteligencia artificial», *El Cronista del Estado Social y Democrático de Derecho*, n.º 100 (2022): 106-115.

Nogueira Alcalá, H. «Aspectos de una Teoría de los Derechos Fundamentales: La Delimitación, Regulación, Garantías y Limitaciones de los Derechos Fundamentales», *Revista Ius et Praxis*, vol. 11, n.º 2 (2005): 15-64.

Palmiotto, F. *The AI Act Roller Coaster: The Evolution of Fundamental Rights Protection in the Legislative Process and the Future of the Regulation*. Cambrigde Core, 2025.

Pérez-Ugena, M. «La inteligencia artificial: definición, regulación y riesgos para los derechos fundamentales», *Estudios de Deusto*, vol. 72, n.º 1 (2024): 307-337.

Sánchez Díaz, M. F. «El impacto de la inteligencia artificial generativa en los derechos humanos», *Revista Eurolatinoamericana de Derecho Administrativo*, vol. 11, n.º 1 (2024).

Thomaidou, A., Limniotis, K. «Navigating Through Human Rights in AI: Exploring the Interplay Between GDPR and Fundamental Rights Impact Assessment», *J. Cybersecur. Priv.*, vol. 5, n.º 1 (2025).

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 60-76

PORTADORES DE DERECHOS FUNDAMENTALES EN EL ENTORNO DE LAS MANIFESTACIONES EMERGENTES DE LA INTELIGENCIA ARTIFICIAL

FUNDAMENTAL RIGHTS HOLDER IN THE ENVIRONMENT OF EMERGING MANIFESTATIONS OF ARTIFICIAL INTELLIGENCE

Federico César Lefranc Weegan

Universidad Marista de Mérida

Resumen

En este momento sería ineludible la responsabilidad de poner a las investigaciones en ciencia y tecnología al servicio de la humanidad, y en ese sentido el análisis acerca de los vínculos entre la ética, la epistemología y los derechos fundamentales que subyacen a las investigaciones, pueden contribuir a proteger la dignidad humana desde el núcleo de los derechos que pudieran ser afectados. Dejar de ignorar el impacto que pueda provocar el uso acrítico de la inteligencia artificial para acercarnos a evaluar su impacto profundo parece urgente.

Palabras clave

Derechos fundamentales, dignidad humana, inteligencia artificial, ética, epistemología

Abstract

At this juncture, it is imperative that we place scientific and technological research at the service of humanity; furthermore, an analysis of the links between ethics, epistemology, and the fundamental rights underlying such research can help protect human dignity by addressing the core rights that may be affected. It seems urgent that we stop ignoring the potential consequences of the uncritical use of artificial intelligence in order to assess its long-term impact.

Keywords

Fundamental rights, human dignity, artificial intelligence, ethics, epistemology

I. Introducción

Las tecnologías, especialmente sus expresiones como la inteligencia artificial, y su conjunción con la dignidad humana requieren de una profunda reflexión. Nunca será fácil abordar el conjunto de características que trazan los límites de la tecnología, así como la usamos en la cotidianidad. La incertidumbre que provoca la falta de una idea homogénea de Inteligencia Artificial, complica las posibles formas de relación que intentamos establecer entre ésta y los llamados derechos fundamentales. Hay que dejar asentado que los principales derechos fundamentales se han reorientado a la protección de la dignidad humana (Häberle, 2003, p. 164 y ss.). Esto ha sido así a partir de la Declaración Universal de los Derechos Humanos de 1948. La dignidad humana, el pensamiento tecnológico, la bioética, y una infinidad de saberes se vinculan en el siglo XXI con la inteligencia artificial y con sus productos en los más variados niveles y complejidades. Se producen constantemente reflexiones sobre estos temas. Ello significa que las reflexiones sobre la dignidad se actualizan continuamente. Uno de los problemas más severos relacionado actualmente con el desarrollo tecnológico, es que, se ha identificado que no está dirigido únicamente a un bienestar colectivo, sino que se le ha enfocado primordialmente como un negocio,

no obstante, las referencias a este bienestar aparecen insistentemente en los discursos mercadológicos.

Claramente uno de los objetivos promovidos por el desarrollo tecnológico es generar ganancias. Esto se logra a partir de un negocio especialmente complejo. O de una serie intrincada de negocios asociados. Reiteramos que en el centro de ese negocio no se sitúan las necesidades más profundas o más humanas de las personas, lo que se busca es cómo aprovechar el mayor número de necesidades de las personas para transformarlas en ganancias. No obstante, sigue habiendo grandes espacios de investigación y desarrollo donde la mirada está puesta en beneficios genuinos para la gente, en cualquiera de los terrenos ya citados.

II. La dignidad humana como sustento

1. Más allá de la concepción dominante

Desde hace algunas décadas la concepción kantiana sobre la dignidad humana ha sido la concepción dominante sobre todo en el ámbito jurídico, pero ésta ha sido criticada con acierto porque expulsa a las mujeres del ámbito de la autonomía, dejándolas fuera con ello, del espacio del ejercicio pleno de la dignidad. Ese aspecto de la crítica pone de relieve los límites del modelo ilustrado que se invoca permanentemente. La crítica es imprescindible para contribuir a la construcción, en diversos ámbitos, de una mirada verdaderamente inclusiva, sobre la dignidad y todo lo que de ella se deriva, en este momento de transformaciones aceleradas en todos los campos de la tecnología. Recordemos que la inteligencia artificial constituye un muy amplio cuerpo de conocimientos que se ha desenvuelto a lo largo ya de aproximadamente siete décadas y lo que se sugiere es que se vea atravesado por una concepción de dignidad humana adecuada a nuestro tiempo (McCarthy, Minsky, Rochester y Shannon, 1955).

Una postura ética se hace sutilmente presente en cada decisión que toma quien hace tecnología al igual que cualquier persona inmiscuida en el desarrollo a profundidad de aquellos conocimientos que a lo largo del tiempo terminan por cobrar relevancia para nuestra sociedad. Así es desde que, en la Modernidad se empezó a dar una estructura sistemática al conocimiento, (Merchant, 2008, p. 732)[1] y es así, actualmente en el ámbito de la inteligencia artificial. Sin embargo, el desarrollador pocas veces se hace consciente de esa postura inicial. Quien trabaja con datos o con hardware muy sofisticado, no siempre se percata de estar asumiéndola. Esta postura inicial impacta en todas las decisiones que se toman para la investigación, en las reflexiones que la acompañan y en los desarrollos que se derivan (Serra, 2015, p. 21). Se menciona esta postura ética inicial, porque es en el terreno de la ética en el que, con mayor fertilidad se ha explorado el tema de la dignidad humana.

1 Tendríamos que recorrer, desde el protagonismo de Bacon, a quien se le atribuye como ingredientes de su investigación metodológica la actitud de inquisidor activo.

2. La importancia de incluir la dignidad en el imaginario de la inteligencia artificial

Lo que identificamos como inteligencia artificial puede ser abordado, a partir de perspectivas como las siguientes: Científica tecnológica, incluida la investigación científica, epistemológica (no cualquier programa de cómputo ni cualquier aparato), económica y de mercado (quiénes los desarrollan, quiénes los controlan), financiera (quiénes se benefician), política, sociológica, ecológico ambiental, antropológica, ética filosófica, o jurídica.

Hay que aclarar de una vez que mucho de lo que se publica acerca de la inteligencia artificial a partir de las ciencias sociales o de las humanidades, se hace a partir del nivel de simples usuarios.

Entonces se describen, reseñan y analizan programas que se asume que utilizan inteligencia artificial, apelando a la verosimilitud de lo que se difunde. Pero la gran mayoría de las personas, incluyendo quien esto escribe, no tienen la preparación especializada que se requiere para distinguir cuándo se encuentran utilizando un sistema de IA y cuándo están frente a variantes de sofisticados programas de cómputo tradicionales. Y si nos enfrentamos a una serie de circuitos, no podríamos diferenciar cuáles están dedicados a la inteligencia artificial y cuáles no. Si el aprendizaje de algún lenguaje de programación relacionado con la inteligencia artificial en realidad provoca un cambio significativo en la manera en la que una persona no especializada puede aprovechar los productos de su campo de conocimientos, todavía no ha sido rigurosamente verificado. Todo lo anterior nos coloca en el limitado rol de usuarios o de consumidores de datos, desconociendo o subestimando la inmensa complejidad humana que nos caracteriza.

Esto no significa que no haya que reconocer el esfuerzo grandísimo de las innumerables personas que hoy dedican su vida y su obra a estos temas. Es claro que los beneficios son incontables en muchas áreas significativas para la sociedad. Sin duda estamos presenciando el desarrollo de tecnologías asombrosas, que paradójicamente nos muestran tan solo la sombra de la complejidad humana. Sin embargo, hay que tomar la distancia suficiente. Por el momento, a quienes escribimos desde las ciencias sociales o desde las humanidades nos hace falta dialogar y compartir reflexiones amplísimas con quienes han desarrollado la capacidad de ir creando y desarrollando los sofisticados sistemas de inteligencia artificial para no perder de vista en aquellos desarrollos ese rasgo esencial que hemos concebido como la dignidad humana.

III. Inteligencia artificial. La elaboración de una narrativa

1. La construcción de un discurso

Términos como: redes neuronales, arquitecturas complejas, procesamiento paralelo, codificación masiva en paralelo y muchos otros tienen, para muchas personas dedicadas a la investigación, únicamente un referente discursivo, pero no pueden ser verificados de primera mano por la mayoría de quienes, en los

campos antes referidos, nos ocupamos del tema. ¿Puede quien esto lee, identificar alguna diferencia entre la inteligencia artificial y otras formas de cómputo? ¿Puede, sin temor a equivocarse, explicar, qué la distingue, o qué no es inteligencia artificial?

Los sistemas de inteligencia artificial pueden ser concebidos como sistemas complejos en los que interactúan necesariamente, aparatos de arquitectura especializada –incluidas sofisticadas interfaces para la adquisición de información - con programas igualmente especializados.

Pero la siguiente idea podría ilustrarnos a partir de un enfoque distinto, podemos concebir la inteligencia artificial primordialmente cómo «la habilidad que tenemos de replicar al cerebro de poder representar las sinapsis, las neuronas, las memorias a corto y largo plazo a través de las matemáticas» (Angulo, 2024)[2].

En un diverso ejercicio de circularidad se puede consultar a un programa de inteligencia artificial. Su autodefinición es la siguiente:

> In summary, the key differences between AI and traditional computing programs lie in their approach to problem-solving and adaptability. AI systems leverage data-driven learning methods to make decisions and perform tasks, while traditional programs rely on predefined instructions and algorithms. (ChatGPT, 2024)[3]

Destacan en esa caracterización el rendimiento y la adaptabilidad al momento de resolver problemas. Y sabemos que la inteligencia artificial se integra como un sistema: equipo especializado al que se agregan programas especializados.

Aclaremos desde este momento que el uso intensivo de la inteligencia artificial no sustituye ni la responsabilidad ni la preparación de las personas, un uso que debería ser, crítico, consciente, responsable y no idealizado.

Por otra parte, cuando se hace investigación sobre la inteligencia artificial desde las ciencias sociales o desde las humanidades, en realidad básicamente lo que se estarán revisando son los discursos que se emiten alrededor de este campo de estudios. Y en el caso de aproximaciones relacionadas con la ética o con la filosofía esto parece adecuado, aunque no suficiente. Una concepción general reciente dirá que:

2 Entrevista con Julio Angulo (2024). Es un ingeniero de investigación y desarrollo especializado en proyectos de inteligencia artificial para una empresa internacional. Él se describe como apasionado por descubrir y explorar nuevas tendencias tecnológicas, como la computación paralela y la IA generativa. En su tiempo libre desconecta de la ingeniería leyendo literatura, practicando yoga y haciendo senderismo en la montaña…

3 «En resumen, las principales diferencias entre la IA y los programas informáticos tradicionales radican en su enfoque para la resolución de problemas y su adaptabilidad. Los sistemas de IA aprovechan los métodos de aprendizaje basados en datos para tomar decisiones y realizar tareas, mientras que los programas tradicionales se basan en instrucciones y algoritmos predefinidos».

> el High-Level Expert Group on sustainable finance (HLEG) de la Comisión Europea definió la IA como aquellos "*sistemas* de software (y posiblemente también de hardware) diseñados por humanos que, ante un objetivo complejo, actúan en la dimensión física o digital: percibiendo su entorno, a través de la adquisición e interpretación de datos estructurados o no estructurados, razonando sobre el conocimiento, procesando la información derivada de estos datos y decidiendo las mejores acciones para lograr el objetivo dado. Los *sistemas* de IA pueden usar reglas simbólicas o aprender un modelo numérico, y también pueden adaptar su comportamiento al analizar cómo el medio ambiente se ve afectado por sus acciones previas". (Cuatrecasas Monforte, 2020, pp. 27-28)[4]

No hay, desde las humanidades o desde las ciencias sociales, una aproximación directa ni al diseño de equipos ni a la programación, ni a la modelación matemática. Se insiste en que la perspectiva dominante será entonces la de los usuarios.

2. Opacidad y otros límites para una comprensión general

El funcionamiento de la inteligencia artificial implica generalmente el procesamiento de datos a partir de capas de algoritmos, y la recirculación de los resultados dando lugar a un aprendizaje automático. Este proceso idealmente requiere transparencia para poder entender, y someter a crítica, la forma en la que han sido obtenidos los resultados. Sin embargo, esta transparencia no siempre es posible. A esa opacidad que se produce durante este complejo proceso, se le conoce también como caja negra.

Una forma de describir el funcionamiento de la IA, es asumirlo como el procesamiento de capas de algoritmos y de sus resultados —que concebimos como aprendizajes—. Estos resultados se van alimentando continuamente conforme se resuelven las diversas capas alcanzándose una gran complejidad en estadios avanzados del procesamiento.

> Incluso han salido nuevas áreas que le llaman ML Ops, machine learning operations, donde es necesario, por ejemplo, reentrenar las redes neuronales,

4 Y, por su parte, en la Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen Normas Armonizadas en materia de Inteligencia Artificial (Ley de Inteligencia Artificial) y se modifican determinados Actos Legislativos de la Unión, publicada el 21 de abril de 2021, se definen los sistemas de IA en su artículo 3.1 como: «el software que se desarrolla empleando una o varias de las técnicas y estrategias que figuran en el anexo I (a saber, Estrategias de aprendizaje automático, incluidos el aprendizaje supervisado, el no supervisado y el realizado por refuerzo, que emplean una amplia variedad de métodos, entre ellos el aprendizaje profundo. Estrategias basadas en la lógica y el conocimiento, especialmente la representación del conocimiento, la programación (lógica) inductiva, las bases de conocimiento, los motores de inferencia y deducción, los sistemas expertos y de razonamiento (simbólico). Estrategias estadísticas, estimación bayesiana, métodos de búsqueda y optimización.) y que puede, para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa».

> estarlas monitoreando si se están como ya no te están dando resultados buenos, entonces hay que hacer un proceso como un feedback. (Angulo, 2024)[5]

Los espacios del tipo caja negra parecen imprescindibles para la inteligencia artificial en acción. La falta de claridad sobre cualquier proceso tiende a implicar un riesgo para el cuidado que le debemos a la dignidad de cada persona. La falta de claridad significa la aceptación a ciegas de decisiones que no sabemos si debemos compartir. En ese terreno la posibilidad de una evaluación ética de los procesos cobra un significado especial.

> Ha surgido toda una corriente dentro de esta disciplina que se orienta a minimizar el impacto de la caja negra de la Inteligencia Artificial. Para ello, busca poner el foco en la investigación en explicabilidad y transparencia de la IA, desarrollando técnicas y métodos que permitan a los expertos comprender mejor cómo funciona esta tecnología. (Barquilla, 2024)

Los alcances de la inteligencia artificial son inimaginables. Esta posibilidad que permite a las computadoras o a las tecnologías asociadas a estas, simular la inteligencia humana y su capacidad de resolver problemas, se alimenta de otros múltiples desarrollos tecnológicos. Desde cuestiones como el reconocimiento facial, hasta la inteligencia artificial generativa, pasando por la modelación del lenguaje natural, la geolocalización, el reconocimiento de voz, los usos especializados en la ciencia y en el diagnóstico médico.

> Today, generative AI can learn and synthesize not just human language but other data types including images, video, software code, and even molecular structures. Applications for AI are growing every day. But as the hype around the use of AI tool in business takes off, conversations around AI ethics and responsible AI become critically important. (Stryker y Kavlakoglu, 2024)[6]

5 «Hay una área que yo todavía no he explorado, una área de la inteligencia artificial que se llama *reinforcement learning*, algo así, se supone que esa área trabaja en eso, que puedan aprender por sí solas, en el (…) *reinforcement learning* es como, como que va aprendiendo con el tiempo, como que se equivoca y aprende, se equivoca y aprende, se equivoca y aprende, se equivoca y aprende, es como su base, el error, la prueba y error» (Angulo, 2024).

6 «Actualmente, la IA generativa puede aprender y sintetizar no solo el lenguaje humano, sino también otros tipos de datos, como imágenes, vídeo, código de software e incluso estructuras moleculares. Las aplicaciones de la IA crecen día a día. Sin embargo, a medida que aumenta el interés por el uso de herramientas de IA en el ámbito empresarial, los debates sobre la ética y la IA responsable se vuelven cruciales.La inteligencia artificial (IA) es una tecnología que permite a las computadoras y máquinas simular la inteligencia humana y la capacidad de resolución de problemas. Por sí sola o combinada con otras tecnologías (por ejemplo, sensores, geolocalización, robótica), la IA puede realizar tareas que, de otro modo, requerirían inteligencia o intervención humana. Los asistentes digitales, la navegación GPS, los vehículos autónomos y las herramientas de IA generativa (como Chat GPT de OpenAI) son solo algunos ejemplos de la IA presente en las noticias y en nuestra vida cotidiana. (…) Las aplicaciones de la IA crecen día a día. Pero a medida

Se habla poco de la manera en la que están involucradas las matemáticas en el pensamiento algorítmico que permite la posibilidad de replicar la naturaleza del cerebro y sus actividades en las máquinas. Y ese es un tercer aspecto que coloca a los no expertos, al margen de la comprensión profunda del tema de la inteligencia artificial (AIMS Mathematics, 2024). La modelación matemática no está al alcance de la mayoría de las personas y conviene no olvidar que Minsky era profesor de matemáticas.

IV. Ética, dignidad humana e inteligencia artificial

1. La dignidad humana como guía permanente

Hay que reconocer abiertamente que en realidad es mucho lo que ni siquiera alcanzamos a ver. Este desconocimiento amplio exige cuando menos prudencia al momento de escribir. Tenemos que reconocer que básicamente lo estamos haciendo sobre un terreno desconocido y que las grandes guías que normalmente utilizamos para orientarnos, en este momento son poco útiles. Es aquí donde todavía deberían jugar un papel relevante nociones como, la moral, la ética y la dignidad humana. «No podemos delegar nuestra responsabilidad a las máquinas. (…) La inteligencia artificial está aquí. Eso significa que hay que ajustarla cada vez más a los valores humanos y a la ética humana» (Zeynep, 2017).

La historia reciente nos debería haber enseñado a ser prudentes. La singularidad humana requiere permanentemente de alguien más, no de una máquina ni de un sistema, sino de una vida que acompañe su fragilidad. Reconocemos que todavía la evaluación de la respuesta de los sistemas de inteligencia artificial, debería de estar a cargo de humanos, que no deberíamos declinar la responsabilidad. No obstante, el imperativo tecnológico nos mostrará que la evaluación también tenderá a delegarse en los propios sistemas (Angulo, 2024)[7].

Hasta aquí, lo que he querido poner de relieve es todo lo que en realidad ignoramos de este vastísimo campo. Por eso habría que revisar desde qué perspectiva nos estamos pronunciando, al momento de escribir o de pretender legislar o normar de algún modo lo que está sucediendo en el terreno de la inteligencia artificial. Sería deseable que se abordara desde una transdisciplina especializada, sería deseable también conservar los pies en la tierra, para interpretar este desconocimiento no como algo apabullante, sino como consecuencia de la vastedad del campo que se pretende abordar. Centrémonos en algunas cuestiones que contribuyan a darle sentido.

que aumenta el interés por el uso de herramientas de IA en los negocios, las conversaciones sobre ética y responsabilidad en la IA se vuelven cruciales» (traducción propia).

7 «Yo digo que sí [que una IA puede evaluar a otra IA], creo que no es lo correcto, pero yo no dudo que haya gente que lo haga, porque para evaluar a una inteligencia artificial pues se necesitan muchos humanos. Hay que pagarles a muchos humanos» (entrevista con Julio Angulo).

2. Limitaciones del enfoque regulatorio. Actualidad de la ley de Hume

La ley de Hume recoge y expone algunas de las grandes limitaciones de las concepciones normativas del mundo. El mundo occidental parece haber heredado del mundo romano la tentación permanente de normar. De darle un orden a todo lo posible, porque el orden transmite seguridad. Y al mismo tiempo, poco caso se le ha hecho a una perspectiva epistemológica fundamental, como la llamada ley de Hume. Esta ley revela un problema epistemológico crucial. La dificultad de verificar rigurosamente la forma de incidir del mundo del deber en el mundo del ser. A pesar de que el Derecho tiene una larga historia, todavía no sabemos si acaso y de qué manera podrían actuar las normas en el mundo de los hechos. La consecuencia de ignorar la ley de Hume, es incurrir en alguna falacia, ya sea normativista al pretender que la promulgación de una norma cambiará la realidad, o naturalista al pretender que la invocación a los hechos puede generar automáticamente un deber. Omitiendo el reconocimiento de que, de premisas de deber solo estamos legitimados a obtener consecuencias de deber.

La pretensión actual de regular hasta el mínimo detalle el desarrollo de los sistemas de inteligencia artificial, no puede dar resultados fácticos por sí misma, aunque la existencia de leyes envía el mensaje de que el tema ha sido atendido. Eso posiblemente sucederá con el Acta de la Comisión Europea (Parlamento Europeo, 2021) y con la reciente y abundante legislación estadounidense (Insight, 2024)[8].

Se revela un inmenso esfuerzo por incidir un terreno que es muy difícil de normar, cuyos procesos y dinámicas, únicamente están al alcance de gente altamente especializada en tecnología, y se desarrollan muy alejados del ámbito

8 AI Watch (2024). "Laws/Regulations directly regulating AI (the "AI Regulations")Currently, there is no comprehensive federal legislation or regulations in the US that regulate the development of AI or specifically prohibit or restrict their use. However, there are existing federal laws that concern AI albeit with limited application. A non-exhaustive list of key examples includes:

Federal Aviation Administration Reauthorization Act, which includes language requiring review of AI in aviation.

- National Defense Authorization Act for Fiscal Year 2019, which directed the Department of Defense to undertake various AI-related activities, including appointing a coordinator to oversee AI activities.
- National AI Initiative Act of 2020, which focused on expanding AI research and development and created the National Artificial Intelligence Initiative Office that is responsible for "overseeing and implementing the US national AI strategy."Nevertheless, various frameworks and guidelines exist to guide the regulation of AI, including:
- The White House Executive Order on AI (titled Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence) which is aimed at numerous sectors, and is premised on the understanding that "[h]arnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks." (…)

The Federal Communications Commission issued a declaratory ruling stating that the restrictions on the use of "artificial or pre-recorded voice" messages in the 1990s era Telephone Consumer Protection Act include AI technologies that generate human voices, demonstrating that regulatory agencies will apply existing law to AI". AI Watch: Global regulatory tracker - United States | White & Case LLP (whitecase.com), consulta junio 4 de 2024

de lo legislable. Recordemos lo siguiente: «Las religiones más difundidas, las leyes más ambiciosas y las doctrinas morales más bienintencionadas, están condenadas al fracaso por su imposibilidad de impactar sobre la realidad técnica» (Lefranc, 2020, p. 123).

Sirva como ejemplo, en el plano de lo atroz, el de las armas letales autónomas respecto de las cuales parecen claramente insuficientes los esfuerzos de la ONU para que los Estados restrinjan a los fabricantes a través de la regulación (ONU, 2023)[9].

La regulación jurídica pertenece a un plano de reflexión específico que es el del deber. Como ya se ha estudiado, no es verificable que el plano del deber tenga algún impacto sobre el plano del ser. En otras palabras, *las leyes no tienen manera de detener las transformaciones tecnológicas*.

Esta aserción, aparentemente simple debería de alertarnos sobre la limitada eficacia de la tremenda avalancha regulatoria que se está desarrollando actualmente. No obstante algunos de los mejores intentos de concretar las garantías materiales que deben de contribuir a proteger la dignidad humana, se han traducido a través de formas diversas de regulación, en prácticas rigurosas a las que han comprometido las grandes corporaciones en materia de inteligencia artificial. La concepción de prácticas es importante porque se presta más claramente al escrutinio y la corrección. Convine estudiar casos como el de la *Guía de Buenas Prácticas del Observatorio del Impacto Social y Ético de la Inteligencia Artificial* (Belloto, 2022).

V. El imperativo tecnológico

1. El imperativo tecnológico y el principio de inevitabilidad

El imperativo tecnológico afirma lo siguiente: «Todo lo que sea posible hacer será hecho, independientemente de nuestra voluntad individual, de nuestra conciencia, de nuestras convicciones éticas, de las religiones o de las leyes existentes» (Lefranc, 2015, p. 277).

Lo que se desprende de este imperativo es que, tenemos la capacidad de desarrollar objetos técnicos, como la inteligencia artificial y esa posibilidad de desarrollarlos la interpretamos como un deber. Por eso poner al descubierto este imperativo que alude específicamente a la ciencia y a la tecnología y que revela una capacidad propia de nuestra especie, debería sosegarnos un poco. Creamos porque somos seres humanos. Somos nosotros, los que tenemos esta posibilidad. Se trata de un imperativo sobre la inevitabilidad, que se está interpretando

9 «Los Sistemas de Armas Autónomos Letales (SAAL) son sistemas de armamento que, mediante el uso de inteligencia artificial (IA), pueden operar en forma autónoma sin intervención humana una vez activados. Debemos actuar ahora para preservar el control humano sobre el uso de la fuerza. Se debe mantener el control humano en las decisiones de vida o muerte. Que las máquinas ataquen de forma autónoma a los humanos es una línea moral que no debemos cruzar», enfatizaron. El derecho internacional debería prohibir las máquinas con el poder y la discreción de acabar con vidas sin intervención humana, señalaron».

como: *ya que podemos hacerlo, debemos desarrollar inteligencia artificial.* Pero uno de los problemas más grandes radica en que lo estamos haciendo sin preguntarnos, quiénes se deben de encargar de estos desarrollos. El principio de inevitabilidad, que se desprende del imperativo tecnológico, permite hasta cierto punto hacer prospectiva. Nos está anticipando que cualquier cosa que se pueda imaginar seriamente, se intentará. Pero intentarlo no está al alcance de todas las personas.

Cuando usamos algunas aplicaciones de inteligencia artificial parece que hay alguien que decide en nuestro nombre, que decide lo que hay que decir y cómo hay que decirlo. No obstante, este potencial nos asombra y, por qué no decirlo, nos disminuye un poco. Va cobrando vida la imagen de un ser humano sometido, receptor pasivo de todo lo que amable pero inevitablemente se le imponga (Cavallé, 2008, p. 245)[10].

> El verdadero peligro, entonces, no es que la gente tome a un chatbot por una persona real; es que comunicarse con los chatbots haga que las personas reales hablen como chatbots, pasando por alto todos los matices y las ironías, diciendo obsesivamente y con precisión lo que creen que quieren decir. (Zizek, 2024)

Cuando nos acomodamos a las reglas de los nuevos sistemas de inteligencia artificial cedemos individualmente algunas de nuestras capacidades representativas. Ya no podemos orientarnos en una ciudad, tampoco redactar un texto profundo o hacer operaciones aritméticas. Estas actividades, que parecen elementales, no lo son, han sido parte de la formación de ese mismo ser humano al que se le ocurrió la posibilidad de una inteligencia artificial. Pero desde hace tiempo, como parte del desarrollo comercial de estos sistemas, se pide al usuario renunciar a muchas de sus habilidades, a cambio de obtener resultados estandarizados. Olvidamos que, el hecho de que algunas de nuestras habilidades puedan ser traducidas a algoritmos, no significa que nosotros, como especie, actuemos a partir de esos mismos algoritmos.

No es pertinente de ninguna manera la comparación constante de la inteligencia artificial —y de los artefactos en general— con lo humano, porque tiende a reducir la infinita complejidad que nos caracteriza, y a mandar un mensaje de subestima, dejando en el olvido el hecho de que siempre se tratará de productos concebidos y desarrollados por el propio ser humano. Los mensajes científicos y mercadotécnicos relacionados con el tema deberían decir algo así: *Juan, María, Arlene y Lucy, desarrollaron un sistema de inteligencia artificial que ha permitido resolver X problema.*

Reconociendo en primer lugar a quienes han logrado un desarrollo tan importante. Pero lo que se suele escribir desaparece a quienes verdaderamente

10 «Detrás de la tecnificación del mundo, de la dictadura de la publicidad y de la información, etc., yace —hace ver Heidegger— lo que quizá sea el rasgo fundamental del actual modo de estar del hombre en el mundo: la búsqueda de seguridad. Una sed de seguridad creciente que, lejos de ser saciada por la organización racional y técnica, es acrecentada por ésta» (Cavallé, 2008, p. 245).

han resuelto la situación, para poner el acento en los productos: *La inteligencia artificial ha resuelto el problema X.*

Este tipo de enunciados brinda erróneamente el carácter de sujeto a un producto del trabajo humano, y no a las personas que lo desarrollaron.

La imagen del ser humano no debería de ser aquella que defina la propia inteligencia artificial, porque estaríamos sometidos a una doble reducción. Primero nos habremos definido en términos manejables para el sistema y luego el sistema volcará la concepción que haya elaborado de nosotros, en términos manejables para sí mismo (Lefranc, 2016, p. 35)[11].

2. Las características que la tecnología hereda

Por último hagamos un resumen de aquellas características de la tecnología que son compartidas por la inteligencia artificial y que no podríamos ignorar precisamente por su impacto en la promoción y garantía de los derechos fundamentales.

2.1 Falibilidad

Los aparatos fallan, cualquier aparato está sujeto a ello. La más sofisticada de las máquinas lo está. Y por supuesto los sistemas de inteligencia artificial no están exentos de fallas, ni en el diseño y funcionamiento físico, ni en el desempeño de los programas. Las fallas por definición no son previsibles, lo que es importante anticipar son las consecuencias más severas que sea posible vislumbrar (Zeynep, 2024)[12]. Ningún sistema de inteligencia artificial se escapa de la posibilidad de fallar. Mientras más delicado sea el trabajo que se está efectuando con un artefacto sofisticado, mayor es la responsabilidad, no únicamente de quien lo opera, sino de quien se encuentre a cargo de la acción en general[13].

11 «La evolución de los pactos, desde la carta magna entre los súbditos y el rey, donde está presente la razón divina, los pactos del mundo moderno, entre los súbditos y el Estado sustentados en la Razón de Estado, hasta llegar al último gran pacto celebrado por la humanidad. El último gran pacto de la humanidad tuvo su origen en una vivencia atroz, la Segunda Guerra Mundial, y se sintetiza en la obligación actual de los Estados de reconocer, proteger y promover la dignidad humana. Los pactos pueden ser explícitos o tácitos, escritos o verbales, más o menos solemnes, deseados o indeseados, pero los pactos están presentes en el siglo XXI, contribuyendo a conformar el imaginario colectivo. Todos estos pactos han tenido sus protagonistas, ¿quién de todos ellos es usted? Ahora, en el inicio del siglo XXI, este último gran pacto parece haber empezado su declive, una vez que se ha manifestado en los hechos la imposibilidad de consolidar un Estado de bienestar generalizado. ¿Necesita un pacto la Sociedad de la Información y del Conocimiento? o ¿Puede funcionar sin él? ¿Entre quiénes se celebraría un pacto así? ¿El pacto en la Sociedad de la Información seguiría siendo inspirado en la dignidad humana?» (Lefranc, 2016, p. 35).

12 «Zeynep Tufekci explica cómo las máquinas inteligentes pueden 'fallar' de maneras que no encajan en los patrones del error humano, de modos que ni esperamos ni predecimos» (Tufekci, 2024).

13 Con Julio Angulo se platicó incluso de aquellos momentos en los que hay que reentrenar las redes neuronales porque ya no te están dando buenos resultados.

2.2 La imposible neutralidad

Hay un discurso permanente sobre la neutralidad de la inteligencia artificial. Se nos dice que no es buena ni mala, sino que depende de para qué se use. Este discurso es erróneo. La tecnología no puede nunca ser neutral. En el caso de la inteligencia artificial hay equipos de personas con formaciones muy especializadas que se dedican ya sea al diseño del *hardware* o al del *software*[14]. Pues bien, estos equipos de personas dedicados al desarrollo de estos complejos sistemas de inteligencia artificial, tienen sus propias historias de vida, su cultura, una ideología, unos valores, y una perspectiva frente a la vida, que invariablemente se reflejan en las decisiones que van tomando en el complejo proceso de diseño, por no hablar de los innumerables sesgos que se va adquiriendo el sistema durante las interacciones con las personas que los entrenan o que los usan.

2.3 Monopolización y dependencia profunda

Todo lo que implica el desarrollo, venta, comercialización, etc. de la inteligencia artificial se encuentra monopolizado —o para no ser tan duros, concentrado en tres o cuatro empresas a nivel mundial—. La dependencia respecto de esas pocas gigantescas corporaciones que concentran los desarrollos es inevitable y todavía no alcanzamos a evaluar sus verdaderas consecuencias. Son muy pocas las grandes compañías, que se dedican a desarrollar sistemas de inteligencia artificial que inciden a nivel mundial (Angulo, 2024)[15]. Muchas de las grandes decisiones que inciden en la economía planetaria se toman en ese entorno.

2.4 Obsolescencias

Los sistemas de inteligencia artificial se están transformando a gran velocidad. Lo mismo el equipamiento que la forma misma de concebir los programas. La obsolescencia se hace presente como en todo desarrollo tecnológico. Lo que hace unos días era apenas una promesa, hoy se ha revelado como inútil. O, por el contrario, temas que fueron subestimados hace algunos años cobran en este momento especial relevancia, como es el caso de la visión artificial (Anymatica, 2024)[16]. Los sistemas caducarán, como parte de su funcionamiento natural, o de

14 Ejemplos de dicho *software* son los siguientes: H2O.ai, Inc., IBM Watsonx, TensorFlow, Vertex AI, Salesforce Inc, Azure Machine Learning Studio, ChatGPT, Google AI, Chatbot, Framer AI, Infosys Nia, Synthesia, Microsoft Azure, Birdeye, Cortana, DataRobot, Inc., Decktopus AI, GitHub Copilot, Otter.ai, SAS, TIDIO.

15 «Lo común es que tu celular mande un request, una llamada a un servidor que está en alguna nube de Asho, de Google, de AWS, allá corre lo que tenga que correr y te regresa la respuesta (…)». Sí, las [tareas] complejas sí, y se concentran en tres empresas. (…) Asho que es de Microsoft, en la nube de Google, GCP se llama, Google y en AWS que es de Amazon. (…) yo creo que el grueso de todo lo que corre en este mundo está concentrado en esas tres empresas, toda nuestra data, la mayoría de nuestra data está ahí» (entrevista con Julio Angulo).

16 Anymatica (2024): «Hoy en día, la visión artificial se sitúa a la vanguardia del avance tecnológico. Los vehículos autónomos utilizan esta tecnología para la navegación vial, los drones para el reconocimiento y la topografía aérea, y el sector sanitario la emplea en el diagnóstico de enfermedades mediante imágenes médicas. El potencial de la visión

forma programada, dando lugar a sustituciones forzosas de tecnología que más de una vez requieren de inversiones inimaginables.

2.5 Vulnerabilidad

Alguien está tratando de *hackear* tu sistema. Es así de manera sistemática y permanente. Atrás de ello puede haber intereses económicos, pero no siempre es así. Precisamente esa cualidad tan humana como es la de poder crear o poder destruir, es la que permite exponer una hipótesis profunda sobre el *hackeo*. En principio se intenta vulnerar los sistemas, como un reto, por el mero hecho de superarlo, por el hecho de demostrar que ninguna barrera es suficiente. Más adelante todo estará servido para los negocios inmensos de la seguridad informática.

2.6 Las limitaciones de la representación digital del mundo

La elaboración de lo digital es sólo una faceta de las capacidades humanas, pero la promoción exhaustiva de los aparatos basados en la inteligencia artificial parecería agotar las opciones. Insistimos en que promover superlativamente esa perspectiva única, reduce peligrosamente la concepción de lo humano y su proyección en el futuro. Podemos aseverar que la digitalización de la mayoría de las cuestiones representables constituye una limitación en la formación de una concepción integral del ser humano. Nuestro cerebro puede operar perfectamente con las representaciones digitalizadas del mundo, pero tiene un potencial muy superior a ello.

2.7 La inevitable circularidad

Mucha de la información que actualmente alimenta los sistemas de inteligencia artificial fue generada previamente. El momento actual parece privilegiado porque estos sistemas pueden aprovechar toda esa información disponible. No obstante irá cobrando importancia la pregunta sobre en qué periodo y de qué formas se fue generando toda esa información. Y qué pasará si decidimos alimentarnos primordialmente de esa información sin reservar espacios de independencia respecto de los sistemas tecnológicos. Si básicamente alimentamos nuestras investigaciones de aquello que se deposita en la red, la posibilidad es que las respuestas elaboradas por la inteligencia artificial, aun siendo sofisticadas, terminarán por ser circulares.

No sobra advertir que nuestros buenos deseos, entre ellos, el de que la tecnología llegue a ser verdaderamente autónoma y alcance la perfección no se le pueden imponer a las características generales de la propia tecnología que, ya advertimos, son compartidas ineludiblemente por los sistemas de inteligencia artificial.

artificial es ilimitado, y la investigación en curso amplía continuamente los límites de lo que la IA puede percibir y comprender» (traducción propia).

VI. Conclusiones. Lo humano

Se va revelando con mayor claridad la necesidad de considerar permanentemente, y a profundidad el tema de la dignidad humana. No reducirlo, ni simplificarlo, tampoco transformarlo en un lema vacío o en una fórmula única. Es vital identificar el carácter polisémico de la expresión, pero no conformarse con ello. Asumir que la dignidad lo mismo se puede inteligir que mostrar, que el reto será darle siempre su valor más hondo individual y colectivamente. Darle ese valor incluso en medio de las discusiones sobre la más intrincada tecnología para que pueda sostenerse como la raíz que da vida y soporte a los derechos fundamentales en ese mismo entorno. Uno de los retos más grandes consiste actualmente en comprender que el contexto ha sufrido profundas transformaciones, que la figura del Estado ha sido desplazada y que, quizás debamos, en ejercicio de nuestra dignidad, aprender a exigir nuevos derechos fundamentales frente nuevas formas de poder.

El campo de la inteligencia artificial recibe constantemente inversiones multimillonarias. Es uno de los ámbitos más dinámicos de la actualidad. El derecho está tratando de anticiparse a esta intensa dinámica. Lo siento, no es ni será posible. Las leyes no pueden dar vida a realidades por venir, su papel es otro.

Da la impresión de que nos encontramos ante procesos inevitables de reducción de los derechos fundamentales ante los aparatos digitales. Pero le corresponde a cada quién refutar esta afirmación y mostrar de cuán variadas formas se pueden ejercer dichos derechos. Es decir, de cuántas inimaginables formas podemos portar los derechos fundamentales, la primera es reconociéndonos en todos los espacios, incluso en el digital, y no sumergiéndonos en el anonimato provocado por la adjudicación omnipresente a la inteligencia artificial. Debemos recordar que la dignidad humana, que nos da identidad, no es un objeto, ni una simple expresión de la que los sistemas se deban apropiar.

La falta de precisión de mucha de la información difundida sobre la propia inteligencia artificial, el hecho de que se trate de un conocimiento muy especializado, pero cuyos secretos se hace parecer que están al alcance de cualquier persona, sumado a la dinámica acelerada con la que se ha ido desarrollando este campo de conocimientos, va provocando una sobrestima difusa de su potencial real.

La consecuencia de esta sobrestima de los artefactos es una subestima de las potencialidades humanas, que busca reducir a cada persona a un limitado proveedor de datos, eso es lo que por ningún motivo deberíamos permitir.

Referencias bibliográficas

AIMS Mathematics. (2024). Special Issue: Advances of Artificial Intelligence-based mathematical modeling optimization in engineering applications. *AIMS Mathematics*.

Anymatica. (2024). AI-Vision Evolution: From Neural Networks to Seeing Machines. *Anymatica*. https://anymatica.com/ai-vision-evolution-from-neural-networks-to-and-seeing-machines

Barquilla, Y. (2024). Qué es la caja negra de la Inteligencia Artificial. *BeeDigital*. https://www.beedigital.es/inteligencia-artificial/que-es-la-caja-negra-de-la-inteligencia-artificial/

Belloto Gómez, J. M. (2022). *GuIA de buenas prácticas en el uso de inteligencia artificial ética*. PWC-OdiseIA. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.pwc.es/es/publicaciones/tecnologia/assets/guia-buenas-practicas-uso-inteligencia-artificial-pwc-odiseia.pdf&ved=2ahUKEwjKiqu1r8OHAxUuIEQIHe9-LJcQFnoECDEQAQ&usg=AOvVaw2hObTyaTw8_FpG_JwJVW9x

Cavallé, M. (2008). *La sabiduría de la no-dualidad. Una reflexión comparada entre Nisargadatta y Heidegger*. Kairos.

Cuatrecasas Monforte, C. (2020). *La Inteligencia Artificial en el proceso penal de instrucción español: posibles beneficios y potenciales riesgos*. Universitat Ramón Llull

Häberle, P. (2003). *El Estado Constitucional*. Universidad Nacional Autónoma de México.

Insight. (2024). AI Watch: Global regulatory tracker - United States. *White & Case*. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states#:~:text=As%2520noted%2520above%252C%2520there%2520is,Responsible%2520innovation%2520and%2520development

Lefranc, F. (2015). Poder de fuego. Acerca de un a violencia sin odio. *El Sistema de Justicia Penal y nuevas formas de observar la cuestión criminal*. Instituto Nacional de Ciencias Penales.

Lefranc, F. (2016). Los sujetos de la SIC. En E. Tellez, *Derecho y TIC. Vertientes actuales*. UNAM.

Lefranc, F. (2020). ¿Debo? Puedo. ¿Puedo? Debo. *Los derechos humanos de los márgenes al centro*. Universidad Autónoma de Tlaxcala.

McCarthy, Minsky, Rochester y Shannon. (1955). *A Proposal For The Dartmouth Summer Research Project On Artificial Intelligence*, http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

Merchant, C. (2008). *"The Violence of Impediments". Francis Bacon and the Origins of Experimentation*. https://warwick.ac.uk/fac/arts/history/students/modules/hi203/group2/bacon.pdf

ONU. (2024). *La ONU y la Cruz Roja urgen a los Estados a poner límites al uso de las armas autónomas*. https://news.un.org/es/story/2023/10/1524637

Parlamento Europeo. (2021). *Artificial intelligence Act*.

Serra, M-Á. (2015). La clave: mejoramiento humano. En A. Cortina y M-Á. Serra, ¿Humanos o Posthumanos? Singularidad tecnológica y mejoramiento humano. Fragmenta Editorial.

Stryker, C. y Kavlakoglu, E. (2024). What is Artificial Intelligence (AI)? *IBM*. https://www.ibm.com/think/topics/artificial-intelligence

Tufekci, Z. (2017). La inteligencia artificial hace que la moral humana sea más importante. *Mujeres con ciencia*. https://mujeresconciencia.com/2017/04/23/

la-inteligencia-artificial-hace-que-la-moral-humana-sea-mas-importante/?fbclid=IwZXh0bgNhZW0CMTEAAR082kj3Ov7jphAffuD5zvG0mWKH16DAgM77HhQHdmOs2Crz8IRCKqBa-So_aem_AeJTcGEwtOTOMKMPq729Qn0CbzmQH-MiyDblpz7jKpQifotQ925Wo2ciwUN7MBjREbnuz6bIRQXBTn7V6WHcJ99

Zizek, S. (2024). Slavoj Zizek sobre la Inteligencia Artificial: "El peligro no es tomar a un chatbot por persona, sino que las personas hablen como chatbots". *Clarín*. https://www.clarin.com/cultura/slavoj-zizek-inteligencia-artificial-peligro-tomar-chatbot-persona-personas-hablen-chatbots-_0_UGO4aDdnqs.html

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 77-89

RESPONSABILIDAD DE LA EMPRESA Y *COMPLIANCE* DIGITAL: EL DEFECTO DE ORGANIZACIÓN EN LA ERA ALGORÍTMICA

CORPORATE RESPONSIBILITY AND DIGITAL COMPLIANCE: THE ORGANIZATIONAL DEFAULT IN THE ALGORITHMIC ERA

José Gregorio Pumarejo Luchón[1]

Estudiante, Universidad Central de Venezuela

1 Abogado egresado de la Universidad Central de Venezuela (UCV). Estudiante en las Especializaciones en Ciencias Penales y Criminológicas y Derecho Procesal en la UCV, en elaboración del Trabajo Especial de Grado en ambas especializaciones. Especialista en Ejercicio de la Función Fiscal de la Escuela de Fiscales del Ministerio Público.

Resumen

La transición hacia una economía globalizada y profundamente digitalizada plantea retos significativos para el Derecho Penal contemporáneo. La responsabilidad penal de las personas jurídicas (RPPJ), fundamentada tradicionalmente en el individualismo y la acción física, evoluciona hacia modelos normativos que sancionan el «defecto de organización». A través de una investigación documental, este artículo analiza cómo el *compliance* digital actúa como el eje central para determinar la responsabilidad de entidades tecnológicas globales, especialmente ante los riesgos derivados de los sistemas algorítmicos. El estudio profundiza en la noción de la empresa como una «fuente de riesgos» que debe ser neutralizada mediante sistemas de autorregulación. Se examina el concepto de *compliance by design*, argumentando que la debida diligencia exige integrar controles preventivos en el código fuente para evitar la «opacidad algorítmica». Se concluye que un programa de cumplimiento eficaz no puede ser meramente formal o cosmético, sino que debe contemplar auditorías técnicas y transparencia proactiva. En la era del algoritmo, la responsabilidad de la empresa surge cuando su estructura técnica, por diseño o negligencia, facilita la comisión de ilícitos informáticos o daños a bienes jurídicos fundamentales.

Palabras clave:

Responsabilidad penal, personas jurídicas, defecto de organización, *compliance* digital, algoritmos, riesgo permitido, ciberseguridad

Abstract

The transition toward a globalized and deeply digitized economy poses significant challenges for contemporary Criminal Law. Corporate Criminal Liability, traditionally based on individualism and physical action, is evolving toward normative models that sanction "organizational failure". Through documentary research, this article analyzes how digital compliance acts as the central axis for determining the liability of global technology entities, especially regarding the risks derived from algorithmic systems. The study delves into the notion of the corporation as a "source of risks" that must be neutralized through self-regulation systems. The concept of compliance by design is examined, arguing that due diligence requires integrating preventive controls into the source code to avoid "algorithmic opacity". It concludes that an effective compliance program cannot be merely formal or cosmetic; it must include technical audits and proactive transparency. In the era of the algorithm, corporate liability arises when its technical structure, whether by design or negligence, facilitates the commission of cybercrimes or damage to fundamental legal interests.

Keywords:

Corporate criminal liability, organizational failure, digital compliance, algorithms, permitted risk, cybersecurity

I. Introducción

En el panorama empresarial contemporáneo, las organizaciones operan en mercados caracterizados por un dinamismo constante y una presión regulatoria creciente. En este entorno, el cumplimiento normativo ha dejado de ser una mera opción operativa para convertirse en un pilar estratégico esencial para evitar sanciones legales y salvaguardar la reputación corporativa[2]. El *corporate compliance* emerge, por tanto, como un sistema integral de autorregulación y gestión de riesgos, diseñado para garantizar que la actividad empresarial se desarrolle bajo estándares éticos y en estricto apego al ordenamiento jurídico vigente[3].

La evolución del Derecho Penal moderno ha desplazado el foco del individualismo tradicional hacia la responsabilidad penal de las personas jurídicas (RPPJ). Bajo esta nueva óptica, la responsabilidad de la empresa no deriva necesariamente de una acción física individual, sino de un «defecto de organización». Como señala la doctrina, las entidades jurídicas pueden ser destinatarias de sanciones cuando su estructura interna facilita o no impide la comisión de ilícitos, evidenciando una falta de control sobre sus recursos y procesos[4]. El cumplimiento normativo actúa aquí como el mecanismo para determinar si la empresa actuó con la debida diligencia, pudiendo funcionar como una eximente o atenuante de la responsabilidad penal[5].

En la era de la transformación tecnológica, este desafío se traslada al ciberespacio, dando lugar al concepto de *compliance digital*. Este se especializa en la gestión de riesgos específicos derivados del uso de tecnologías de la información, tales como los delitos informáticos, la protección de datos y la seguridad de los activos digitales. Ante la sofisticación de la criminalidad digital, el legislador y la doctrina han reconocido la necesidad de establecer deberes de vigilancia sobre fuentes de peligro técnico, como los sistemas algorítmicos, cuya falta de supervisión puede generar daños a gran escala[6].

II. El fundamento de la responsabilidad: el defecto de organización

La doctrina penal contemporánea ha superado la visión antropomórfica que impedía sancionar a las corporaciones. Hoy se admite que la responsabilidad de la empresa no nace de una acción física aislada, sino de una estructura

2 Castagnino, D. T., «Una aproximación al concepto de Corporate Compliance», *Revista Venezolana de Derecho Mercantil*, n.º 3 (2019), p. 102.

3 Vaudo Godina, L., «Compliance corporativo como política de prevención de actos que perjudican la reputación organizacional», *Revista Venezolana de Derecho Mercantil*, n.º 8 (2022), p. 165.

4 Santacruz Salazar, A. Y., *La responsabilidad penal de las personas jurídicas en Venezuela y sus consecuencias socio-económicas* (Caracas: Universidad Central de Venezuela, 2016), p. 45.

5 Pérez Dupuy, M. I., «Función del Compliance en la legislación penal venezolana», *Revista CENIPEC*, n.º 34 (2022), p. 238.

6 Rivero Núñez, E. N., «Responsabilidad de las personas jurídicas ante la comisión de delitos informáticos», *Revista Derecho y Tecnología*, n.º 5 (2019), p. 52.

organizacional que, por diseño o negligencia, facilita la comisión de ilícitos. Este fenómeno se denomina «defecto de organización» o *culpa in vigilando* institucional.

Como sostiene Silvina Bacigalupo, la capacidad de acción de la persona jurídica debe entenderse desde una perspectiva institucional; la empresa «actúa» a través de sus órganos y de la gestión de sus procesos internos[7]. Esta visión desplaza el enfoque desde el individuo hacia el sistema, donde la responsabilidad surge cuando el accionar lesivo es producto del consenso de sus órganos de decisión, el uso de sus recursos o la falta de supervisión adecuada sobre los mismos[8].

Santiago Mir Puig refuerza esta tesis al señalar que la pena a la persona jurídica no se basa en una culpabilidad psicológica (imposible de verificar en un ente ideal), sino en la peligrosidad objetiva e instrumental de la organización[9]. La empresa es considerada una «fuente de riesgos» que el Derecho debe neutralizar mediante la exigencia de una reorganización interna.

La Infracción del Deber de Control: Enrique Bacigalupo enfatiza que el fundamento de la sanción radica en la omisión de medidas de vigilancia que eran exigibles a la administración[10]. En este sentido, la responsabilidad penal de las personas jurídicas (RPPJ) no es una responsabilidad objetiva o por el hecho de otro, sino una responsabilidad por el "hecho propio": el no haber estructurado una organización capaz de prevenir delitos[11].

Un sistema de autorregulación idóneo —el *compliance*— sirve de referencia normativa para afirmar que la empresa actúa dentro del «riesgo permitido». Si la organización implementa controles proporcionales a su peligrosidad, se excluye la imputación objetiva del resultado lesivo, pues el derecho no exige la eliminación total del riesgo, sino su gestión diligente[12]. De acuerdo con Judith Méndez, esta autorregulación permite alinear el comportamiento empresarial con los valores de ética e integridad, reduciendo drásticamente la tasa de criminalidad corporativa[13].

III. El algoritmo como fuente de riesgo en el *compliance* digital

En el ecosistema de las plataformas tecnológicas de alcance global, el algoritmo no es un elemento neutro, sino una herramienta de gestión dirigida a la maximización de beneficios y la retención del usuario (*engagement*). No obstante,

7 Bacigalupo Saggese, S., *La responsabilidad penal de las personas jurídicas: Un estudio sobre el sujeto del Derecho Penal*, tesis doctoral (Madrid: Universidad Autónoma de Madrid, 1997), p. 85.

8 Rivero Núñez, «Responsabilidad de las personas jurídicas…», p. 55.

9 Mir Puig, S., *Derecho Penal. Parte General*, 10.ª ed. (Barcelona, Reppertor, 2016), p. 182.

10 Bacigalupo, E., *Derecho penal: Parte general*, 2.ª ed. (Buenos Aires: Hammurabi, 1999), p. 270.

11 Santacruz Salazar, *La responsabilidad penal*…, p. 48.

12 Pérez Dupuy, «Función del Compliance…», p. 242.

13 Méndez, J., «Análisis del criminal compliance como mecanismo de regulación de la responsabilidad penal de las personas jurídicas», *Iurídica*, vol. 8, n.º 1 (2024), p. 39.

cuando este diseño prioriza la viralización de contenidos nocivos o «retos» peligrosos sin una supervisión adecuada, se transforma en una fuente de riesgo jurídicamente desaprobada que compromete la responsabilidad de la corporación.

Ante la sofisticación de la criminalidad tecnológica, es imperativo superar el enfoque individualista. La doctrina señala que la responsabilidad de estos entes surge cuando el accionar lesivo es producto de la falta de supervisión sobre sus propios recursos tecnológicos[14]. En el ámbito de los delitos informáticos, el legislador venezolano ha acogido la tesis de que la impunidad en organizaciones complejas solo se evita mediante la atribución de responsabilidad penal a la persona jurídica cuando esta se beneficia de la falta de control sobre sus sistemas[15].

El *compliance* digital exige que la empresa identifique al algoritmo como un foco de peligro. Bajo la óptica de la gestión de riesgos, un programa de cumplimiento efectivo debe contemplar medidas de vigilancia y control proporcionales a la magnitud del riesgo gestionado[16]. En consecuencia, una red social que carece de mecanismos de interrupción inmediata frente a conductas autolesivas o ilícitas está incurriendo en una quiebra de sus deberes de cuidado, revelando un programa «cosmético» o ineficaz[17].

El cumplimiento normativo no puede limitarse a manuales estáticos; debe fomentar una cultura que permee la estructura de toma de decisiones, incluyendo el equipo de desarrollo técnico[18]. El concepto de *compliance by design* implica que la prevención de riesgos penales debe integrarse desde la fase de programación. Como indica Alvear-Tobar, la cultura de cumplimiento busca alinear el proceder de la empresa con estándares éticos para evitar que la búsqueda del lucro ignore la protección de bienes jurídicos fundamentales[19]. En este contexto, si una empresa detecta patrones de riesgo en su algoritmo y no realiza los ajustes técnicos necesarios para mitigarlos, incurre en una responsabilidad por omisión de sus deberes de control institucional[20].

IV. El *compliance* como eximente o atenuante

La implementación de programas de cumplimiento normativo trasciende la esfera ética para consolidarse como un elemento de relevancia típica y punitiva. En el marco de la responsabilidad penal de las personas jurídicas (RPPJ), la idoneidad de estos programas constituye el baremo principal para evaluar la diligencia organizacional.

Los programas de *compliance* no deben interpretarse únicamente como herramientas privadas, sino como verdaderos deberes empresariales de colaboración

14 Rivero Núñez, «Responsabilidad de las personas jurídicas…», p. 58.

15 Rivero Núñez, «Responsabilidad de las personas jurídicas…», p. 61.

16 Pérez Dupuy, «Función del Compliance…», p. 248.

17 Pérez Dupuy, «Función del Compliance…», p. 249.

18 Alvear-Tobar, E. J., «El compliance y la cultura de cumplimiento como mecanismos para prevenir la corrupción», *593 Digital Publisher*, vol. 9, n.º 6-1 (2024), p. 128.

19 Méndez, «Análisis del criminal compliance…», p. 40.

20 Mir Puig, *Derecho Penal. Parte General*, p. 815.

con el Estado en la lucha contra la criminalidad corporativa[21]. Como señala Pérez Dupuy, el cumplimiento normativo actúa como un sistema de gestión de riesgos penales diseñado para prevenir la comisión de delitos en el seno de la empresa, alineando el interés privado con el orden público[22].

La autorregulación empresarial incide directamente en la determinación de la pena. Un programa de cumplimiento «idóneo» -capaz de detectar y neutralizar el riesgo antes de su cristalización en un resultado delictivo- puede funcionar como una eximente o, al menos, como una atenuante muy cualificada[23]. Según Judith Méndez, esta capacidad reguladora reduce la intervención del *ius puniendi* cuando la empresa demuestra haber agotado sus deberes de cuidado[24].

La efectividad de un programa de cumplimiento se mide por su proporcionalidad respecto al riesgo gestionado. En el entorno digital, donde los daños pueden ser masivos e inmediatos, la vigilancia no puede ser estática. Como sostiene Rivero Núñez, ante los avances de la criminalidad informática, la empresa debe ejercer un control real sobre sus recursos tecnológicos[25], lo que implica mecanismos de moderación en tiempo real y ajuste constante de algoritmos. La ausencia de estas medidas revela un cumplimiento «cosmético», incapaz de desvirtuar el defecto de organización[26].

1. Ámbito crítico: el *compliance* digital frente a la «opacidad algorítmica»

Un programa de cumplimiento que se limita a la superficie normativa, sin auditar la lógica subyacente de sus procesos automatizados, es ineficaz. En la era algorítmica, el «fraude de ley» se manifiesta en la configuración de sistemas que priorizan resultados económicos (viralidad, tráfico de datos) sobre la seguridad jurídica[27].

Desde la teoría de sistemas (Basabe Serrano), la empresa es un sistema complejo que procesa información; si los canales internos ignoran los riesgos digitales, el sistema presenta un defecto de organización[28]. La «opacidad algorítmica» no es una excusa para la impunidad, sino la prueba de una organización defectuosa que ha perdido el control sobre sus decisiones[29].

21 Bacigalupo, *Derecho penal: Parte general*, p. 272.

22 Pérez Dupuy, «Función del Compliance…», p. 240.

23 Santacruz Salazar, *La responsabilidad penal…*, p. 52.

24 Méndez, «Análisis del criminal compliance…», p. 45.

25 Rivero Núñez, «Responsabilidad de las personas jurídicas…», p. 55.

26 Pérez Dupuy, «Función del Compliance…», p. 249.

27 Modolell González, J. L., *Persona jurídica y responsabilidad penal* (Caracas, Universidad Central de Venezuela, 2002), p. 35.

28 Basabe Serrano, S., *Responsabilidad penal de las personas jurídicas desde la teoría de sistemas* (Quito: Corporación Editora Nacional, 2003), p. 88.

29 Basabe Serrano, *Responsabilidad penal…*, p. 92.

Rafael Chanjan Documet señala que el «hecho de conexión» entre el injusto individual y la cultura corporativa se halla en el diseño[30]. Si la organización fomenta un entorno sin supervisión ética de los desarrolladores, se produce una ruptura de la responsabilidad. Muchas empresas adoptan modelos de autorresponsabilidad pura que, en la práctica, funcionan como una «vicarización» encubierta, imputando fallos técnicos al azar[31].

Por su parte, Juan Luis Modolell sostiene que la empresa es penalmente responsable si «pudo y debió» programar de otra manera[32]. Un algoritmo que facilita delitos informáticos no es un ente autónomo, sino un «equivalente funcional» de la voluntad de la administración que omitió los filtros de seguridad necesarios[33] [34].

La autorregulación digital busca definir el umbral del «riesgo permitido», pero en sectores de alta tecnología este umbral se ha desplazado unilateralmente por las corporaciones. Un *compliance* idóneo no es el que cumple con un *due diligence* formal, sino el que demuestra una prevención activa y constante[35]. Sin auditoría externa de algoritmos y transparencia proactiva, el *compliance* digital se convierte en una herramienta de impunidad[36].

Como sostiene Silvina Bacigalupo, la responsabilidad no puede diluirse en la complejidad técnica; la empresa es responsable de las decisiones «automatizadas» que emanan de su estructura[37]. El riesgo aquí es la creación de programas de cumplimiento «cosméticos» o *paper compliance*, donde la empresa posee manuales extensos pero sus sistemas técnicos operan bajo una lógica de beneficio que ignora deliberadamente la prevención del delito informático. La crítica radica en que, sin una auditoría externa de algoritmos y una transparencia proactiva, el *compliance* digital se convierte en una herramienta de impunidad corporativa en lugar de un mecanismo de control[38].

Siguiendo a Mir Puig, la sanción a la persona jurídica no tiene un fin retributivo, sino preventivo-especial: su objetivo es obligar a la corporación a reorganizarse para que no vuelva a delinquir[39]. El *compliance* es la prueba de que dicha reorganización ha ocurrido. Sin embargo, en el ámbito digital, esta reorganización

30 Chanjan Documet, R. H., *La responsabilidad penal de las personas jurídicas* (Lima: PUCP, 2015), p. 11.

31 Chanjan Documet, *La responsabilidad penal*…, p. 12.

32 Modolell González, *Persona jurídica y responsabilidad penal*, p. 48.

33 Boldova Pasamar, M. Á., «Teoría del delito y persona jurídica», *Revista Peruana de Ciencias Penales*, n.º 35 (2021), p. 65.

34 Boldova Pasamar, «Teoría del delito y persona jurídica», p. 72.

35 Palao Vizcardo, E., «La responsabilidad penal de las personas jurídicas en las operaciones de M&A: ¿Se puede eximir a través de un adecuado due diligence?», *THEMIS, Revista de Derecho*, n.º 84 (2023), p. 315.

36 Santacruz Salazar, *La responsabilidad penal*…, p. 55.

37 Bacigalupo Saggese, *La responsabilidad penal*…, p. 112.

38 Pérez Dupuy, «Función del Compliance…», p. 249.

39 Mir Puig, *Derecho Penal. Parte General*, p. 815.

debe ser técnica: si el algoritmo no se reprograma para filtrar el riesgo, la «reorganización» es inexistente y el defecto de organización persiste.

2. La inversión de la carga de la prueba y el *compliance by design*

En entornos de alta complejidad tecnológica, la empresa debe probar su diligencia más allá del mero cumplimiento formal de estándares. No basta con tener un *compliance officer*; es necesario aplicar el principio de *compliance by design*, integrando controles preventivos en el código fuente desde su concepción[40]. De lo contrario, el programa de cumplimiento carece de la «idoneidad» necesaria y se reduce a una estrategia de defensa procesal[41].

Para que el *compliance* digital opere como eximente, se requieren auditorías independientes que certifiquen que los sistemas de IA no están sesgados hacia riesgos no permitidos[42]. Si la empresa se ampara en el secreto comercial para no revelar el funcionamiento de su algoritmo tras un evento lesivo, el tribunal debería interpretar dicha opacidad como un indicio de defecto de organización[43].

La transición del *compliance* reactivo al proactivo exige que las tecnológicas globales utilicen su propia tecnología para predecir y bloquear patrones de riesgo. La negligencia en la actualización de estos sistemas predictivos podría asimilarse, en términos de política criminal, a una forma de imprudencia grave o de incumplimiento intencional de deberes de supervisión (lo que algunos autores, como Silva Sánchez, denominan riesgo normativamente desaprobado con conocimiento de las circunstancias). Si bien resulta discutible trasladar la categoría del dolo eventual —de raigambre individual y psicológica— al ente colectivo, ciertos pronunciamientos jurisprudenciales y una parte de la doctrina (Gómez-Jara Díez) admiten un dolo corporativo funcional cuando la alta dirección conoce la alta probabilidad de producción de ilícitos y no adopta medidas. No obstante, se trata de una posición minoritaria y sometida a críticas fundadas[44].

V. La dimensión técnica del *compliance*: de *paper compliance* a *algorithmic accountability*

El mayor riesgo del *compliance* en la era digital es su degradación hacia modelos meramente formales. En organizaciones de alta complejidad tecnológica, el defecto de organización no se encuentra en la ausencia de manuales de conducta, sino en la arquitectura del *software*.

40 Méndez, «Análisis del criminal compliance…», p. 52.

41 Pérez Dupuy, «Función del Compliance…», p. 251.

42 Alvear-Tobar, «El compliance y la cultura de cumplimiento…», p. 130.

43 Santacruz Salazar, *La responsabilidad penal…*, p. 64.

44 *Vid*. Carbonell Mateu, J. C., *Derecho penal: parte general* (Valencia, *Tirant lo Blanch*, 2019), pp. 512-515, quien rechaza la construcción del dolo corporativo por analogía *in malam partem*.

1. El sesgo algorítmico como defecto de organización inducido

El sesgo algorítmico no es un mero error técnico fortuito, sino una manifestación de la voluntad corporativa y un fallo estructural en el control de riesgos. Si un algoritmo de recomendación es entrenado con datos que incentivan conductas de riesgo (trastornos alimenticios, retos de asfixia) para maximizar el lucro, la empresa no está ante un «accidente», sino configurando un sistema de peligrosidad objetiva[45].

Como sostiene Silvina Bacigalupo, la persona jurídica debe responder por la «programación de su propia impunidad»[46]. Si el sistema corporativo procesa estímulos que producen una salida lesiva sin activar mecanismos de autoprotección, desde una perspectiva normativa y funcional (Basabe Serrano) podría sostenerse que el sistema presenta un déficit de organización equivalente a un incumplimiento grave de los deberes de vigilancia. No obstante, una corriente doctrinal relevante (Zaffaroni, Schünemann) rechaza atribuir culpabilidad al ente colectivo en sentido estricto, prefiriendo hablar de responsabilidad objetiva mitigada por la posibilidad de exoneración mediante programas de cumplimiento idóneos[47].

Siguiendo a Rivero Núñez, la responsabilidad penal surge cuando el accionar es producto del consenso de los órganos de decisión[48]. En la era digital, este consenso se materializa en métricas de éxito (KPI) que priorizan el *engagement* sobre la seguridad, anulando la cultura de cumplimiento real[49].

Para Modolell, el sesgo algorítmico es un «hecho propio» de la organización porque nace de sus recursos y su gestión, no de un acto ajeno del programador[50]. Como indica Chanjan Documet, el defecto de organización se evidencia cuando el sistema de control no opera eficazmente sobre los activos tecnológicos[51].

Si la empresa oculta el sesgo bajo el manto del secreto comercial, impide la supervisión estatal y refuerza su peligrosidad. Un programa de cumplimiento que no someta el código a auditorías éticas recurrentes constituye un defecto de organización inducido para evadir la sanción penal[52] [53].

45 Bacigalupo Saggese, *La responsabilidad penal*…, p. 145.

46 Bacigalupo Saggese, *La responsabilidad penal*…, p. 148.

47 *Vid*. Zaffaroni, E. R., *Tratado de Derecho Penal*, t. V (Buenos Aires: Ediar, 2012), pp. 456-460, donde se niega la culpabilidad de la persona jurídica y se propone un sistema de consecuencias accesorias administrativas.

48 Rivero Núñez, «Responsabilidad de las personas jurídicas…», p. 61.

49 Rivero Núñez, «Responsabilidad de las personas jurídicas…», p. 61.

50 Pérez Dupuy, «Función del Compliance…», p. 249.

51 Modolell González, *Persona jurídica y responsabilidad penal*, p. 48.

52 Chanjan Documet, *La responsabilidad penal*…, p. 12.

53 Boldova Pasamar, «Teoría del delito y persona jurídica», p. 65.

2. Auditorías de código y ciberseguridad como deberes de vigilancia

El *compliance officer* debe evolucionar hacia un perfil híbrido técnico-jurídico, capaz de supervisar auditorías de código fuente y protocolos de ciberseguridad. La falta de mecanismos técnicos para la detección temprana de anomalías (bots, cuentas automatizadas, contenidos delictivos) constituye una quiebra de los deberes de control institucional[54]. Desde la teoría de sistemas, el cumplimiento digital es un subsistema de control que permite al sistema organizacional mantener su integridad[55]. Si la organización carece de auditorías de código, priva a su «sistema nervioso» de los sensores necesarios para identificar riesgos[56].

La idoneidad de un programa de cumplimiento es directamente proporcional a la capacidad económica y tecnológica del ente[57]. Una corporación global que lidera el desarrollo de IA no puede alegar «imposibilidad técnica» ante fallos en su moderación de contenidos. La omisión de filtros preventivos se traduce en culpa organizacional[58]. En el ámbito digital, esta evitabilidad se traduce en la auditoría algorítmica.

La «evitabilidad del hecho» (Modolell) se traduce en la auditoría algorítmica. Si una auditoría de código hubiera revelado que el sistema priorizaba contenidos de odio o violencia, y la empresa no aplicó los parches por razones de rentabilidad, incurre en un dolo de omisión organizacional[59]. El deber de vigilancia hoy no se agota en la supervisión humana, sino en la validación constante de la seguridad del *software*[60].

Para Boldova Pasamar, un protocolo de ciberseguridad ineficaz es el equivalente funcional de una negligencia grave en una fábrica industrial[61]. Como sostiene Alvear-Tobar, la verdadera cultura de cumplimiento se mide en la capacidad de anticiparse a la brecha de seguridad mediante inversión proactiva en defensa técnica[62].

En definitiva, la responsabilidad penal de la persona jurídica en la era digital no busca frenar la innovación, sino garantizar que esta se desarrolle dentro de los límites del riesgo permitido. Las organizaciones que omitan controles proporcionales a su capacidad técnica estarán aceptando -bajo la figura del dolo eventual organizacional- las consecuencias penales derivadas de su estructura defectuosa.

El Derecho Penal moderno, al abordar la responsabilidad penal de las personas jurídicas (RPPJ), exige que la intensidad del control sea equivalente a la intensidad del riesgo generado. Como señala Pérez Dupuy, el cumplimiento

54 Boldova Pasamar, «Teoría del delito y persona jurídica», p. 72.

55 Basabe Serrano, *Responsabilidad penal…*, p. 105.

56 Pérez Dupuy, «Función del Compliance…», p. 248.

57 Chanjan Documet, *La responsabilidad penal…*, p. 12.

58 Modolell González, *Persona jurídica y responsabilidad penal*, p. 50.

59 Rivero Núñez, «Responsabilidad de las personas jurídicas…», p. 58.

60 Boldova Pasamar, «Teoría del delito y persona jurídica», p. 72.

61 Alvear-Tobar, «El compliance y la cultura de cumplimiento…», p. 129.

62 Pérez Dupuy, «Función del Compliance…», p. 248.

normativo es un deber de colaboración con el Estado que se intensifica en función de la estructura de la empresa[63]. Para una plataforma digital que gestiona millones de interacciones diarias, el estándar de "debida diligencia" no se satisface con una moderación humana reactiva, sino que exige la implementación de filtros de IA preventiva de última generación.

Desde la perspectiva de Juan Luis Modolell, la sanción penal a la empresa se justifica en la medida en que esta «pudo y debió» organizarse para evitar el hecho[64]. En el contexto de las tecnológicas, la «posibilidad de evitar» está directamente vinculada a su vanguardia técnica. Si la empresa tiene la capacidad de programar algoritmos predictivos para optimizar sus ingresos, tiene, por correlación lógica y jurídica, la capacidad de programar algoritmos para detectar y bloquear la comisión de delitos informáticos o la viralización de contenidos lesivos. La omisión de esta capacidad técnica se traduce en un defecto de organización por infracción del deber de vigilancia, elevando el umbral de reproche penal[65].

Siguiendo la tesis sistémica de Santiago Basabe Serrano, la empresa digital debe entenderse como un sistema autopoiético que define sus propias reglas de funcionamiento a través del código[66]. Si el sistema decide, por una política de maximización de beneficios, no asignar recursos técnicos a la supervisión del cumplimiento, está incurriendo en una «autoprogramación de la culpabilidad». La empresa no es una víctima de la complejidad de su propia tecnología, sino la arquitecta de un entorno de impunidad.

Como advierte Rafael Chanjan Documet, el «hecho de conexión» entre el injusto y la corporación se perfecciona cuando la cultura corporativa tolera o fomenta la ausencia de controles[67]. En las grandes tecnológicas, esta tolerancia se manifiesta en la opacidad de los algoritmos de moderación. La «caja negra» deja de ser un secreto comercial para convertirse en un indicio de culpabilidad cuando se utiliza para ocultar la ineficacia de los programas de cumplimiento.

VI. Conclusiones

Del análisis realizado se derivan las siguientes conclusiones, que se presentan asumiendo explícitamente las controversias dogmáticas existentes.

El defecto de organización como fundamento predominante de la RPPJ en entornos digitales. La mayoría de la doctrina que admite la responsabilidad penal de la persona jurídica sitúa su fundamento no en una acción física aislada, sino en una estructura organizacional que, por diseño o negligencia, facilita la comisión de ilícitos. En la era algorítmica, dicho defecto se manifiesta cuando la empresa omite deberes de vigilancia sobre sus propias fuentes de peligro (los algoritmos). No obstante, una corriente minoritaria (Zaffaroni, Schünemann)

63 Pérez Dupuy, «Función del Compliance…», p. 240.

64 Modolell González, *Persona jurídica y responsabilidad penal*, p. 50.

65 Rivero Núñez, «Responsabilidad de las personas jurídicas…», p. 58.

66 Basabe Serrano, *Responsabilidad penal…*, p. 115.

67 Chanjan Documet, *La responsabilidad penal…*, p. 12.

niega toda forma de RPPJ y propone trasladar esta materia al Derecho administrativo sancionador.

El *compliance* digital como exigencia jurídica, pero con distintos modelos de imputación. El cumplimiento normativo deja de ser una opción corporativa para convertirse en una necesidad jurídica, ya sea como causa de exoneración en los sistemas de RPPJ o como atenuante en modelos de responsabilidad objetiva mitigada. Un programa de cumplimiento eficaz debe ser proporcional a la magnitud de la operación y a la capacidad técnica del ente, y no puede limitarse a un mero formalismo (*paper compliance*).

La falta de auditorías externas e independientes de algoritmos, así como el amparo en el secreto comercial para no revelar el funcionamiento de sistemas lesivos, puede interpretarse como un indicio cualificado de defecto de organización. Sin embargo, no debe confundirse con una presunción iuris et de iure de culpabilidad, ni trasladar mecánicamente al ente colectivo categorías subjetivas como el dolo. El trabajo aboga por un enfoque basado en el incumplimiento grave de deberes objetivos de supervisión, más que en una ficticia «voluntad corporativa».

Necesidad del *compliance by design* y de auditorías técnicas, sin incurrir en analogías problemáticas. La prevención de riesgos penales debe integrarse en el código fuente desde la fase de programación. Las auditorías de código y la ciberseguridad proactiva constituyen deberes de vigilancia ineludibles para las plataformas tecnológicas globales. Ahora bien, calificar su omisión como «dolo eventual corporativo» resulta controvertido; por ello, se propone describir esos supuestos como imprudencia organizativa grave o incumplimiento consciente de deberes de supervisión, dejando abierto el debate dogmático.

La «imposibilidad técnica» no es admisible para las grandes tecnológicas. Aquellas empresas que disponen de recursos y vanguardia técnica para maximizar beneficios también poseen la capacidad de implementar filtros preventivos. La omisión de estos filtros constituye, para la doctrina dominante, el núcleo del injusto que fundamenta la sanción penal a la persona jurídica (como déficit de organización), mientras que para las posturas negacionistas debería dar lugar a sanciones administrativas o consecuencias accesorias.

El *compliance* digital como evolución necesaria, pero con cautelas dogmáticas. No basta con la prevención humana; se requiere una prevención algorítmica auditable y transparente. Sin embargo, al construir la RPPJ en entornos digitales, es preciso evitar trasladar acríticamente categorías del Derecho Penal individual (culpabilidad, dolo) a la persona jurídica. El trabajo asume una posición intermedia: admite la RPPJ como herramienta legítima de política criminal, pero modula sus fundamentos en clave de riesgo organizativo y deberes de supervisión, sin necesidad de recurrir a una culpabilidad corporativa en sentido estricto.

En suma, la responsabilidad penal de la empresa en la era del algoritmo no busca frenar la innovación, sino garantizar que esta se desarrolle dentro de los límites del riesgo permitido. Las organizaciones que omitan controles proporcionales a su capacidad técnica estarán asumiendo —al menos desde la perspectiva de la prevención general positiva— las consecuencias jurídicas derivadas de su

estructura defectuosa. El cumplimiento es, por tanto, la última ratio de la integridad empresarial en un mundo donde el código, efectivamente, se ha convertido en ley.

Referencias bibliográficas

Alvear-Tobar, E. J. «El compliance y la cultura de cumplimiento como mecanismos para prevenir la corrupción». *593 Digital Publisher*, vol. 9, n.º 6-1 (2024): 124-133.

Basabe Serrano, S. *Responsabilidad penal de las personas jurídicas desde la teoría de sistemas*. Quito: Corporación Editora Nacional, 2003.

Boldova Pasamar, M. Á. «Teoría del delito y persona jurídica». *Revista Peruana de Ciencias Penales*, n.º 35 (2021): 57-80.

Carbonell Mateu, J. C. *Derecho penal: parte general*. Valencia: Tirant lo Blanch, 2019.

Chanjan Documet, R. H. *La responsabilidad penal de las personas jurídicas*. Lima: PUCP, 2015.

Gómez-Jara Díez, C. *La culpabilidad penal de la empresa*. Madrid: Marcial Pons, 2005.

Heine, G. *La responsabilidad penal de las empresas*, trad. J. L. González. Valencia: Tirant lo Blanch, 1996.

Méndez, J. «Análisis del criminal compliance como mecanismo de regulación de la responsabilidad penal de las personas jurídicas». *Iurídica*, vol. 8, n.º 1 (2024): 37-60.

Mir Puig, S. *Derecho Penal. Parte General*, 10.ª ed. Barcelona: Reppertor, 2016.

Modolell González, J. L., *Persona jurídica y responsabilidad penal*. Caracas: Universidad Central de Venezuela, 2002.

Pérez Dupuy, M. I. «Función del Compliance en la legislación penal venezolana». *Revista CENIPEC*, n.º 34 (2022): 235-260.

Rivero Núñez, E. N. «Responsabilidad de las personas jurídicas ante la comisión de delitos informáticos», *Revista Derecho y Tecnología*, n.º 5 (2019): 47-61.

Santacruz Salazar, A. Y. *La responsabilidad penal de las personas jurídicas en Venezuela y sus consecuencias socio-económicas*. Caracas: Universidad Central de Venezuela, 2016.

Schünemann, B. «La responsabilidad penal de las empresas en el Derecho alemán y europeo». *Revista de Derecho Penal y Criminología*, n.º 14 (2004): 31-58.

Silva Sánchez, J. M. *La expansión del Derecho penal. Aspectos de la política criminal en las sociedades postindustriales*, 2.ª ed. Madrid: Civitas, 2001.

Vaudo Godina, L. «Compliance corporativo como política de prevención de actos que perjudican la reputación organizacional». *Revista Venezolana de Derecho Mercantil*, n.º 8 (2022): 163-182.

Zaffaroni, E. R., Alagia, A. y Slokar, A. *Derecho Penal. Parte General*, 2.ª ed. Buenos Aires: Ediar, 2002.

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 90-107

CONTRATOS DE USO DE INTELIGENCIA ARTIFICIAL GENERATIVA Y LEY APLICABLE: LÍMITES DEL REGLAMENTO ROMA I EN EL ENTORNO DIGITAL

CONTRACTS FOR THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE AND THE APPLICABLE LAW: THE LIMITS OF THE ROME I REGULATION IN THE DIGITAL ENVIRONMENT

Carlos Domínguez Padilla

Universidad de Cádiz

Resumen

El presente trabajo analiza los contratos de uso de sistemas de inteligencia artificial generativa desde la perspectiva del derecho internacional privado. Se sostiene que estas relaciones contractuales no se ajustan plenamente a las categorías tradicionales del Reglamento (CE) n.º 593/2008 (Roma I), en la medida en que no se estructuran en torno a una prestación delimitada, sino al acceso a una capacidad generativa caracterizada por su opacidad y parcial indeterminación. Aunque su calificación como contratos de prestación de servicios permite aplicar el artículo 4.1 b) del Reglamento, dicha solución resulta funcionalmente insuficiente debido a la disociación entre la localización del prestador y el lugar en que se despliegan los efectos del contrato. El trabajo pone de relieve que la determinación de la ley aplicable ya no puede entenderse de forma aislada, al encontrarse la relación contractual estructuralmente condicionada por normas de aplicación necesaria del derecho de la Unión Europea, en particular el Reglamento de Inteligencia Artificial y el Reglamento General de Protección de Datos. Asimismo, se examina la articulación entre responsabilidad contractual y extracontractual, destacando las dificultades de imputación del daño y la creciente centralidad de la gestión del riesgo.

Palabras clave

Inteligencia artificial generativa, Reglamento Roma I, normas de aplicación necesaria, ley aplicable, responsabilidad extracontractual

Abstract

This article examines contracts for the use of generative artificial intelligence systems from the perspective of private international law. It argues that such contracts do not fully fit within the traditional categories underlying Regulation (EC) No. 593/2008 (Rome I), as they are not structured around a clearly defined performance, but rather around access to a generative capability characterized by opacity and partial indeterminacy. Although their qualification as service contracts allows for the application of Article 4(1)(b) Rome I, this solution proves functionally insufficient due to the dissociation between the service provider's location and the place where the contractual effects materialize. The analysis shows that the determination of the applicable law can no longer be understood in isolation, as the contractual relationship is structurally shaped by overriding mandatory provisions of EU law, including the AI Act and the General Data Protection Regulation. The paper also addresses the interaction between contractual and non-contractual liability, highlighting the difficulties of attributing harm and the growing importance of risk management in this field.

Keywords:

Generative artificial intelligence, Rome I Regulation, overriding mandatory provisions, applicable law, non-contractual liability

I. La irrupción de la inteligencia artificial generativa y la crisis de las categorías contractuales en el derecho internacional privado

La incorporación de sistemas de inteligencia artificial (IA) generativa en la práctica no se limita a introducir mejoras técnicas o incrementos de eficiencia. Implica, en realidad, una alteración más profunda de la forma en que se estructuran determinadas relaciones jurídicas, especialmente en el ámbito contractual. En este sentido, la generalización de modelos de uso general, que son accesibles mediante interfaces digitales y capaces de generar contenidos complejos a partir de instrucciones abiertas, está dando lugar a formas de vinculación que no encajan sin fricciones en las categorías clásicas del derecho de obligaciones.

Desde la perspectiva del derecho internacional privado, este fenómeno resulta particularmente problemático. En particular, el Reglamento (CE) n.º 593/2008 (Roma I) presupone, en su lógica interna, que es posible identificar con cierta claridad el tipo contractual de que se trata. Sobre esa base opera la determinación de la ley aplicable. Sin embargo, cuando el contrato presenta una configuración opaca o poco nítida, la propia operación de calificación deja de ser un paso técnico previo para convertirse en un auténtico problema de fondo.

Los contratos de uso de sistemas de IA generativa, habitualmente formalizados mediante condiciones generales asociadas a plataformas digitales o interfaces API, constituyen un ejemplo paradigmático de ello. No se trata, en sentido estricto, de contratos de prestación de servicios tal como han sido tradicionalmente entendidos, pero tampoco pueden identificarse sin más con licencias de *software* ni con contratos de suministro digital. Lo que se contrata no es la ejecución de una actividad concreta ni la cesión de un programa delimitado, sino el acceso a una infraestructura algorítmica cuyo funcionamiento efectivo permanece, en buena medida, fuera del alcance del usuario[1].

Aquí reside una de las claves del problema: el objeto del contrato no es un resultado ni una actividad, sino la posibilidad de generar resultados. El proveedor no se compromete a producir un *output* determinado, sino a habilitar un entorno en el que el usuario, a partir de sus propios *inputs*, obtiene respuestas cuya configuración concreta no puede ser anticipada con precisión. Esta indeterminación no es accidental ni marginal, sino que forma parte de la propia lógica de estos sistemas y tiene consecuencias jurídicas evidentes en materia de obligaciones, distribución de riesgos e imputación de resultados.

A ello se suma una asimetría informativa especialmente intensa, pues el proveedor controla, al menos en términos estructurales, el diseño y entrenamiento del modelo, mientras que el usuario se limita a interactuar con una interfaz cuya lógica interna desconoce. No se trata simplemente de una desigualdad informativa más o menos típica, sino de una opacidad técnica que condiciona directamente la posibilidad de aplicar categorías jurídicas tradicionales como el error, la diligencia o el incumplimiento. Por ello, en este contexto, los instrumentos conceptuales clásicos empiezan a mostrar sus límites.

1 Muñoz Vela, J. M., «Inteligencia artificial generativa. Desafíos para la propiedad intelectual», *Revista de Derecho de la UNED*, n.º 33 (2024): 17-75.

Tampoco puede ignorarse que la relación contractual desborda el esquema bilateral. El uso de sistemas de IA generativa se inserta en un entorno tecnológico más amplio, en el que intervienen múltiples actores y cuya operatividad depende de infraestructuras complejas. Además, los efectos del uso del sistema no se agotan en la relación entre proveedor y usuario, sino que pueden proyectarse sobre terceros, lo que introduce una dimensión adicional que el esquema contractual tradicional no capta adecuadamente.

Todo ello tiene una consecuencia directa desde el punto de vista conflictual. La determinación de la ley aplicable no puede plantearse exclusivamente como una operación de subsunción en las categorías del artículo 4 del Reglamento Roma I[2]. Aunque, en una primera aproximación, estos contratos puedan reconducirse a la noción de prestación de servicios, dicha calificación resulta insuficiente para reflejar la realidad jurídica de la relación. En particular, el criterio de la residencia habitual del prestador no siempre coincide con el lugar en el que se concentran los efectos del contrato, lo que evidencia la insuficiencia del criterio clásico de conexión.

De hecho, en la mayoría de los supuestos relevantes, el proveedor se encuentra establecido fuera de la Unión Europea, mientras que el uso del sistema y, por tanto, la generación de efectos jurídicos relevantes, se produce dentro de ella. Esta disociación entre localización formal y material no es un dato accesorio, sino uno de los rasgos definitorios de estos contratos, y pone de manifiesto una tensión difícil de resolver dentro del esquema clásico del Reglamento Roma I.

A partir de aquí, la cuestión deja de ser meramente técnica porque no se trata solo de determinar qué ley resulta aplicable al contrato, sino de reconocer que esa ley ya no opera de manera aislada. En los contratos de uso de IA generativa, la *lex contractus* se ve necesariamente condicionada por un conjunto de normas de aplicación necesaria y otras disposiciones imperativas que responden a intereses estructurales del ordenamiento europeo[3] y que se proyectan sobre la relación con independencia de la ley elegida o designada por las partes.

Esta idea es la que articula el presente trabajo, partiendo de la premisa de que los contratos de uso general de IA generativa no solo plantean problemas de encaje en las categorías tradicionales, sino que evidencian un desplazamiento en la forma en que el derecho internacional privado ordena las relaciones contractuales en el entorno digital. La ley aplicable sigue siendo relevante, pero ya no basta por sí sola.

Sobre esta base, el análisis se centrará, en primer lugar, en los problemas de calificación jurídica que plantean estos contratos; en segundo lugar, en los límites de los criterios de conexión del Reglamento Roma I para captar su realidad funcional; y, finalmente, en el papel que desempeñan las normas de aplicación

2 Merchán Murillo, A., «Ley aplicable en la contratación automatizada», *Derecho internacional privado en acción: algunas cuestiones: IV Foro europeo de Derecho internacional privado*, coord. A. Fernández Pérez, 2024, pp. 71-94.

3 Deskoski, T y Dokovski, V., «Lex Contractus for Specific Contracts under the Rome I Regulation», *Iustinianus Primus L. Rev.*, vol. 10 (2019), p. 1.

necesaria del Derecho de la Unión Europea y el régimen de responsabilidad en la configuración del marco jurídico aplicable.

II. La calificación jurídica de los contratos de uso de inteligencia artificial generativa y sus límites en el marco del Reglamento Roma I

La determinación de la naturaleza jurídica de los contratos de uso general de IA generativa constituye, probablemente, el primer obstáculo relevante desde la perspectiva del derecho internacional privado, en la medida en que condiciona de forma directa la aplicación del método conflictual previsto en el Reglamento (CE) n.º 593/2008 (Roma I). Como es sabido, la lógica interna del Reglamento descansa sobre una previa operación de calificación jurídica del contrato, que permite reconducir el supuesto a alguna de las categorías contempladas en su artículo 4 o, en su defecto, acudir a la cláusula de escape del artículo 4.3. Sin embargo, cuando el contrato presenta una configuración funcional híbrida, como sucede en los contratos analizados en este trabajo, dicha operación deja de ser un ejercicio meramente técnico para convertirse en un auténtico problema estructural[4].

En una primera aproximación, la tendencia dominante consiste en calificar estos contratos como contratos de prestación de servicios digitales. Esta solución responde a una lógica funcional inmediata: el proveedor pone a disposición del usuario una herramienta operativa accesible a distancia, mediante la cual este puede obtener resultados a partir de sus propias interacciones con el sistema. Desde esta perspectiva, la prestación del proveedor consistiría en garantizar el acceso, disponibilidad y correcto funcionamiento del servicio, lo que permitiría subsumir el contrato en la categoría del artículo 4.1 b) del Reglamento Roma I[5].

No obstante, esta calificación, aun siendo operativa, resulta claramente insuficiente. A diferencia de los contratos de prestación de servicios tradicionales, en los que la actividad del prestador puede ser identificada y evaluada en términos relativamente precisos, los sistemas de IA generativa introducen una dimensión cualitativamente distinta. El proveedor no se limita a ejecutar una actividad conforme a parámetros previamente definidos, sino que habilita el acceso a un sistema cuya conducta se basa en modelos estadísticos complejos y cuya respuesta concreta depende de múltiples variables, muchas de ellas inaccesibles para el usuario. En consecuencia, la prestación no puede describirse adecuadamente como una actividad determinada, sino como la puesta a disposición de una capacidad generativa.

Esta precisión no es meramente terminológica, sino jurídicamente decisiva. En los contratos de servicios clásicos, la obligación del prestador se articula en torno a la diligencia en la ejecución de una actividad, lo que permite valorar su cumplimiento o incumplimiento con base en criterios relativamente estables. En

4 Merchán Murillo, A., «Ley aplicable en caso de responsabilidad a las plataformas digitales», *Plataformas digitales: aspectos jurídicos*, dir. A. Martínez Nadal, 2021, pp. 483-500.

5 Castelló Pastor, J. J., Guerrero Pérez, A. y Martínez Pérez, M., *Derecho de la contratación electrónica y comercio electrónico en la Unión Europea y en España* (Tirant lo Blanch, 2021).

cambio, en los contratos de uso de IA generativa, la relación entre la conducta del proveedor y el resultado obtenido se encuentra mediada por un sistema autónomo cuya lógica interna no es plenamente controlable ni por el proveedor ni por el usuario. Ello introduce un elemento de indeterminación estructural que dificulta la aplicación de las categorías tradicionales de incumplimiento contractual.

A esta complejidad se añade que estos contratos presentan, en muchos casos, rasgos propios de otras figuras jurídicas. Así, desde una perspectiva material, pueden identificarse elementos próximos a las licencias de uso de *software*, en la medida en que el usuario accede a una herramienta tecnológica cuya utilización queda sujeta a determinadas condiciones. Sin embargo, esta analogía resulta igualmente insuficiente. A diferencia de las licencias clásicas, en las que el objeto del contrato es un programa informático delimitado y transferido, aunque sea de forma no exclusiva, en los contratos de IA generativa no existe una cesión del *software* en sentido estricto, sino un acceso remoto a una infraestructura dinámica, susceptible de modificación continua sin intervención del usuario.

Del mismo modo, podrían apreciarse ciertos paralelismos con los contratos de suministro de contenidos o servicios digitales, tal como han sido configurados en el derecho de la Unión Europea. Sin embargo, esta categoría tampoco permite captar adecuadamente el fenómeno analizado. Mientras que en los contratos de suministro digital el contenido o servicio objeto del contrato puede ser identificado con relativa precisión, en los sistemas de IA generativa el resultado no está predeterminado, sino que se construye en cada interacción a partir de un proceso de generación que combina los *inputs* del usuario con la estructura interna del modelo.

La dificultad para encajar estos contratos en las categorías tradicionales pone de manifiesto una tensión más profunda en el sistema del Reglamento Roma I. El método conflictual clásico presupone la existencia de tipos contractuales relativamente estables, a los que pueden asociarse criterios de conexión predeterminados. Sin embargo, la evolución tecnológica ha dado lugar a formas contractuales que no responden a estas tipologías cerradas, sino que se configuran como estructuras abiertas, en las que confluyen elementos de diversa naturaleza.

En este contexto, la calificación como contrato de prestación de servicios debe entenderse como una solución pragmática orientada a preservar la operatividad del sistema conflictual, pero no como una descripción adecuada de la realidad jurídica subyacente. Esta aproximación permite aplicar el criterio de la residencia habitual del prestador y facilita la determinación de la ley aplicable, pero lo hace a costa de simplificar en exceso la naturaleza de la relación contractual[6].

La cuestión se complica aún más si se tiene en cuenta la frecuente inclusión de cláusulas de elección de ley en las condiciones generales que rigen estos contratos. En la práctica, los grandes proveedores de sistemas de IA generativa imponen la aplicación de ordenamientos jurídicos concretos, habitualmente los de determinados Estados de los Estados Unidos, en ejercicio de la autonomía de

6 Merchán Murillo, A., *Derecho internacional privado y nuevas tecnologías* (Colex, 2026), pp. 108-111.

la voluntad reconocida en el artículo 3 del Reglamento Roma I. Esta circunstancia desplaza parcialmente la relevancia de la calificación como mecanismo de determinación de la ley aplicable, pero no elimina los problemas derivados de la interacción entre la ley elegida y las normas de aplicación necesaria del foro.

En efecto, la elección de ley no opera en estos supuestos de manera absoluta; el artículo 3 del Reglamento Roma I reconoce la primacía de la autonomía de la voluntad, pero esta se encuentra limitada por la aplicación de normas imperativas, en particular cuando concurren elementos que sitúan la relación contractual en el ámbito de protección del Derecho de la Unión Europea[7]. Así, incluso en presencia de una cláusula de elección de ley a favor de un ordenamiento extracomunitario, el contrato seguirá sometido a las disposiciones de aplicación necesaria del derecho de la UE cuando sus efectos se proyecten sobre su territorio[8].

Desde esta perspectiva, la calificación no pierde relevancia, sino que adquiere una nueva dimensión, pues no se trata únicamente de determinar la categoría contractual a efectos de aplicar el artículo 4 del Reglamento Roma I, sino de identificar la función económica del contrato y su inserción en el ecosistema normativo europeo[9]. Esta aproximación funcional permite comprender por qué la simple subsunción en la categoría de prestación de servicios resulta insuficiente para resolver las cuestiones jurídicas que plantean estos contratos.

En última instancia, la dificultad de calificación de los contratos de uso de inteligencia artificial generativa refleja una transformación más amplia en la estructura del derecho internacional privado, pues la creciente complejidad de las relaciones jurídicas en el entorno digital cuestiona la adecuación de las categorías tradicionales y exige una reinterpretación de los criterios de conexión a la luz de las nuevas realidades tecnológicas. En este sentido, la calificación como contrato de prestación de servicios debe entenderse como un punto de partida, pero no como una solución definitiva.

III. La determinación de la ley aplicable en defecto de elección y la relativización de la *lex contractus*

La determinación de la ley aplicable a los contratos de uso general de IA generativa constituye el verdadero punto de inflexión del análisis desde la perspectiva del derecho internacional privado, en la medida en que pone de relieve, de forma especialmente intensa, las tensiones existentes entre el método conflictual clásico y la realidad económica y tecnológica sobre la que debe proyectarse. En efecto, si bien la calificación de estos contratos como contratos de prestación

7 De Miguel Asensio, P. A., «Reglamento (UE) 2024/1689 de Inteligencia Artificial y Derecho Internacional Privado», *Revista española de derecho internacional*, vol. 77, n.º 1 (2025): 223-236.

8 De Miguel Asensio, P. A, «Cláusulas de elección del derecho aplicable», *Cláusulas en los contratos internacionales: redacción y análisis*, dir. S. Sánchez Lorenzo, 2021, pp. 316-344.

9 De Miguel Asensio, P. A., «Contratación comercial internacional», *Derecho de los negocios internacionales*, coord. J. C. Fernández Rozas, R. Arenas García, P. A. de Miguel Asensio, 2024, pp. 293-404.

de servicios permite, en principio, activar el criterio de conexión previsto en el artículo 4.1 b) del Reglamento Roma I, la aplicación de dicha norma plantea dificultades que trascienden el plano meramente técnico para situarse en el ámbito de la coherencia estructural del sistema[10].

Conforme al citado precepto, en defecto de elección de ley por las partes, el contrato de prestación de servicios se regirá por la ley del país en que el prestador tenga su residencia habitual. Aplicado a los contratos de uso de IA generativa, este criterio conduce, en la mayoría de los casos, a la designación de ordenamientos jurídicos extracomunitarios, habida cuenta de que los principales proveedores de estos servicios se encuentran establecidos fuera de la Unión Europea, fundamentalmente en los Estados Unidos. Desde un punto de vista formal, la solución resulta clara: el contrato se rige por la ley del Estado de establecimiento del proveedor, en cuanto parte que realiza la prestación característica.

Sin embargo, esta aparente claridad se atenúa cuando se examinan los efectos reales del contrato y su inserción en el espacio jurídico europeo. A diferencia de otros contratos de prestación de servicios en los que la actividad del prestador presenta una conexión territorial más identificable, los servicios de IA generativa se caracterizan por una profunda disociación entre el lugar de establecimiento del proveedor y el lugar en el que se producen sus efectos. El acceso al sistema se realiza a través de infraestructuras digitales globales, mientras que los resultados generados se despliegan de manera inmediata en el entorno en el que se encuentra el usuario.

Esta disociación introduce una tensión significativa en la aplicación del artículo 4.1 b) del Reglamento Roma I. El criterio de la residencia habitual del prestador responde a una lógica de previsibilidad y seguridad jurídica, en la medida en que permite identificar, de forma relativamente sencilla, el ordenamiento aplicable. No obstante, en el contexto de los servicios digitales basados en inteligencia artificial, dicha lógica entra en fricción con la realidad económica de la relación, en la que el centro de gravedad del contrato no se sitúa necesariamente en el Estado del proveedor, sino en el entorno en el que el servicio es efectivamente utilizado y despliega sus efectos.

Esta circunstancia invita a cuestionar si la aplicación automática del artículo 4.1 b) resulta adecuada para captar la conexión más significativa del contrato. El propio Reglamento Roma I prevé, en su artículo 4.3, la posibilidad de apartarse de la regla general cuando del conjunto de circunstancias resulte que el contrato presenta vínculos manifiestamente más estrechos con otro país. Sin embargo, esta cláusula de escape presenta un carácter excepcional y corrector, y su activación exige una evidencia clara de la inadecuación del criterio general[11]. Su utilización sistemática en el ámbito de los contratos de inteligencia artificial generativa podría, además, comprometer los objetivos de previsibilidad y seguridad jurídica que inspiran el Reglamento.

10 Merchán Murillo, *Derecho internacional privado y nuevas tecnologías*, pp. 123-128.

11 Ortega Giménez, A., «Artificial intelligence, smart contracts and private international law in Spain», *Los retos que plantea la inteligencia artificial a la administración pública*, coord. L. S. Heredia Sánchez, 2025, pp. 249-282.

En la práctica, la tensión entre la conexión formal y la conexión material no suele resolverse mediante la aplicación del artículo 4.3, sino a través de la intervención de normas de aplicación necesaria en el sentido del artículo 9 del Reglamento Roma I, que operan con independencia de la ley designada conforme al método conflictual[12]. De este modo, aunque el contrato se rija formalmente por la ley del Estado del proveedor, su ejecución queda condicionada por un conjunto de disposiciones imperativas del Derecho de la Unión Europea que responden a la protección de intereses públicos esenciales.

El Reglamento (UE) 2024/1689 sobre inteligencia artificial constituye, en este sentido, un ejemplo paradigmático. Su ámbito de aplicación territorial no se limita a los proveedores establecidos en la UE, sino que se extiende a todos aquellos sistemas de inteligencia artificial que se comercialicen, pongan en servicio o utilicen en el territorio de la UE, con independencia del lugar de establecimiento del proveedor[13]. Esta proyección extraterritorial implica que los contratos de uso de inteligencia artificial generativa, aunque formalmente sometidos a un derecho extracomunitario, deban respetar las obligaciones impuestas por el Reglamento cuando sus efectos se despliegan en la UE[14].

Una lógica similar se observa en el Reglamento General de Protección de Datos, cuyo artículo 3 establece un criterio de aplicación territorial basado en el lugar en que se encuentran los interesados cuyos datos son tratados o en el que se ofrecen bienes o servicios[15]. En consecuencia, el tratamiento de datos personales realizado en el contexto de un contrato de uso de inteligencia artificial generativa puede quedar sometido al RGPD, incluso cuando el responsable del tratamiento se encuentre fuera de la UE y el contrato se rija por una ley extranjera.

La coexistencia de la ley designada conforme al Reglamento Roma I y de las normas de aplicación necesaria del Derecho de la Unión Europea da lugar a un fenómeno de fragmentación normativa especialmente intenso en estos contratos digitales contemporáneos. La relación contractual deja de estar regida por un único ordenamiento jurídico para quedar sujeta a una pluralidad de normas que operan en distintos niveles y responden a lógicas de aplicación diferenciadas. Esta fragmentación no constituye una anomalía, sino una manifestación estructural de la complejidad del espacio jurídico internacional.

En este contexto, la función del artículo 4.1 b) del Reglamento Roma I se ve necesariamente relativizada. Si bien sigue desempeñando un papel relevante en la determinación de la ley aplicable al contrato en sentido estricto, su capacidad para ordenar de manera completa la relación jurídica se encuentra limitada por la intervención de normas imperativas que responden a finalidades regulatorias

12 Calvo Caravaca, A. L., «El Reglamento Roma I Sobre la Ley Aplicable a las Obligaciones Contractuales: Cuestiones Escogidas», *Cuadernos de derecho transnacional*, vol. 1, n.º 2 (2009), pp. 52-133.

13 De Miguel Asensio, «Reglamento (UE) 2024/1689…», pp. 223-236.

14 Ortega Giménez, A., «El impacto extraterritorial del Reglamento europeo de inteligencia artificial», *Anuario de la Facultad de Derecho*, n.º 17 (2024): 221-239.

15 De Miguel Asensio, P. A., «Reglamento General de Protección de Datos: Coordinación entre la aplicación pública y la aplicación privada», *La Ley Unión Europea*, n.º 112 (2023).

específicas. La ley designada conforme al método conflictual deja así de ser el único marco de referencia para la relación contractual, para integrarse en un sistema normativo más amplio.

Esta transformación plantea cuestiones de gran calado sobre la adecuación del modelo conflictual clásico a las nuevas realidades tecnológicas. El Reglamento Roma I fue concebido en un contexto caracterizado por una mayor territorialidad de las relaciones contractuales y una menor densidad normativa supranacional[16]. Sin embargo, en el entorno digital actual, marcado por la deslocalización funcional de las actividades y la proliferación de normas con vocación extraterritorial, la determinación de la ley aplicable ya no puede considerarse una operación suficiente para resolver, por sí sola, los problemas jurídicos que plantea la relación contractual.

Los contratos de uso de IA generativa ponen así de manifiesto los límites del artículo 4.1 b) del Reglamento Roma I como criterio exclusivo de conexión. Sin cuestionar su validez ni su utilidad, resulta necesario reconocer que su aplicación debe ser complementada por un análisis más amplio, capaz de tomar en consideración la interacción entre la ley designada y las normas de aplicación necesaria del Derecho de la UE. Solo desde esta perspectiva puede alcanzarse una comprensión adecuada del régimen jurídico aplicable a estos contratos y garantizarse una protección efectiva de los intereses en juego en el ecosistema digital contemporáneo.

IV. Normas de aplicación necesaria, orden público y su proyección en los contratos de inteligencia artificial generativa

La creciente relevancia de las normas de aplicación necesaria en el ámbito de los contratos de uso de IA generativa obliga a replantear, en términos más precisos, la función que estas desempeñan dentro del sistema del Reglamento Roma I. Tradicionalmente, el artículo 9 ha sido concebido como una cláusula de carácter excepcional, destinada a permitir la aplicación de determinadas disposiciones imperativas del foro en supuestos en los que concurren intereses públicos especialmente intensos. Sin embargo, en el contexto de la economía digital, y en particular en relación con los sistemas de IA generativa, su función práctica parece haber experimentado una expansión significativa que trasciende su utilización meramente puntual.

En efecto, la elevada densidad normativa que caracteriza al Derecho de la UE en materia digital ha dado lugar a un conjunto de disposiciones que no solo inciden de manera aislada en la relación contractual, sino que condicionan de forma estructural su desarrollo y ejecución. Reglamentos como el Reglamento General de Protección de Datos, el Reglamento de Inteligencia Artificial, el Reglamento de Servicios Digitales o el Reglamento de Datos configuran un entramado normativo que opera con independencia de la ley aplicable al contrato y que responde a la protección de intereses públicos esenciales vinculados

16 Calvo Caravaca, A. L. y Carrascosa González, J., *Tratado crítico de Derecho Internacional Privado. Vol. VI. Derecho de los Negocios Internacionales II* (Edisofer Libros Jurídicos, 2024), p. 493.

tanto al funcionamiento del mercado digital como a la salvaguarda de derechos fundamentales[17].

Desde esta perspectiva, la aplicación de estas normas no puede ser entendida como una mera corrección puntual del régimen contractual, sino como la manifestación de un marco normativo que opera de forma concurrente con la ley designada conforme al método conflictual. El contrato deja así de ser un espacio regido de manera exclusiva por la *lex contractus* para convertirse en un punto de intersección entre distintos niveles normativos, cada uno de los cuales responde a una lógica propia.

El artículo 9 del Reglamento Roma I proporciona el fundamento técnico para esta intervención, al permitir la aplicación de disposiciones cuya observancia se considera esencial para la salvaguarda de intereses públicos. No obstante, la utilización de esta categoría en el ámbito digital presenta particularidades relevantes. A diferencia de las normas clásicas de policía, tradicionalmente vinculadas a sectores como el control de cambios, la protección del consumidor o el Derecho laboral, las disposiciones del Derecho de la Unión Europea en materia digital no se limitan a regular aspectos concretos de la relación jurídica, sino que establecen marcos regulatorios amplios que condicionan de manera global el funcionamiento de determinadas actividades[18].

Así, el Reglamento General de Protección de Datos configura un sistema integral que abarca desde la licitud del tratamiento hasta la responsabilidad del responsable, incluyendo los derechos de los interesados y los mecanismos de control. De manera análoga, el Reglamento de Inteligencia Artificial introduce un conjunto de obligaciones que afectan al diseño, desarrollo y utilización de los sistemas de inteligencia artificial, incidiendo directamente en la forma en que se prestan los servicios objeto del contrato.

La consecuencia de esta configuración es que la aplicación de estas normas no puede considerarse meramente excepcional, sino que adquiere un carácter estructural en la configuración del régimen jurídico aplicable. En otras palabras, el contrato de uso de IA generativa se encuentra, desde su origen, inserto en un marco normativo que delimita su contenido y condiciona su ejecución, con independencia de la ley designada conforme al método conflictual.

Esta evolución no exige necesariamente una revisión del concepto de normas de aplicación necesaria, pero sí invita a reconsiderar su función en el ámbito digital. Más que como una excepción al sistema, estas normas operan, en la práctica, como un elemento constante en la configuración de las relaciones contractuales en entornos tecnológicamente avanzados.

En este contexto, la autonomía de la voluntad se ve necesariamente relativizada. Aunque las partes pueden elegir la ley aplicable al contrato conforme al

17 Ortega Giménez, A., «Evolución tecnológica, protección de datos de carácter personal y derecho internacional privado», *Derecho internacional privado en acción: algunas cuestiones: IV Foro europeo de Derecho internacional privado*, coord. A. Fernández Pérez, 2024, pp. 49-64.

18 Domínguez Padilla, C., «Leyes de policía en la contratación con sistemas de inteligencia artificial», *Anuario Español de Derecho Internacional Privado*, n.º 24 (2024): 107-129.

artículo 3 del Reglamento Roma I, dicha elección se encuentra limitada por la obligación de respetar las disposiciones imperativas del derecho de la UE cuando el contrato presenta una conexión suficiente con su territorio. Este límite no debe ser entendido como una restricción arbitraria de la libertad contractual, sino como una consecuencia inherente a la necesidad de garantizar la eficacia de las políticas públicas en el entorno digital[19].

Por su parte, la noción de orden público mantiene su función como cláusula de cierre del sistema, permitiendo excluir la aplicación de una ley extranjera cuando su resultado sea manifiestamente incompatible con los principios fundamentales del foro. No obstante, en el ámbito de IA generativa, la protección de derechos fundamentales como pueden ser la dignidad de la persona, la no discriminación o la transparencia, se articula principalmente a través de normas específicas de carácter imperativo, más que mediante la invocación directa del artículo 21 del Reglamento Roma I.

En este sentido, puede sostenerse que el conjunto de normas del derecho digital de la UE refleja, de manera indirecta, los valores fundamentales del ordenamiento de la UE en el ámbito tecnológico. Sin embargo, más que como una categoría jurídica autónoma, esta idea debe entenderse como una clave interpretativa que permite explicar la coherencia de la intervención normativa europea en este sector.

Desde una perspectiva práctica, la proyección de estas normas plantea importantes implicaciones para la configuración contractual. Las cláusulas que pretendan excluir o limitar la responsabilidad del proveedor, restringir los derechos del usuario o eludir las obligaciones derivadas de la normativa europea pueden resultar ineficaces en la medida en que contravengan disposiciones de carácter imperativo[20]. Asimismo, el incumplimiento de estas obligaciones puede dar lugar a la imposición de sanciones administrativas, con independencia de la ley aplicable al contrato.

La intervención de las normas de aplicación necesaria en los contratos de uso de IA generativa no constituye un fenómeno marginal, sino un elemento central en la configuración de su régimen jurídico. Estas normas no se limitan a corregir el resultado del método conflictual en supuestos excepcionales, sino que operan como un marco estructural que condiciona de manera decisiva la relación contractual. Esta realidad obliga a replantear el papel del Reglamento Roma I en el entorno digital contemporáneo y a reconocer que la determinación de la ley aplicable, aunque sigue siendo una operación necesaria, ya no resulta suficiente por sí sola para comprender la complejidad del régimen jurídico aplicable a estos contratos.

19 Jacquet, J. M., «La aplicación de las leyes de policía en materia de contratos internacionales», *Anuario español de Derecho internacional privado*, vol. 10 (2010): 35-48.

20 Merchán Murillo, *Derecho internacional privado y nuevas tecnologías*, pp. 118-123.

V. Responsabilidad derivada del uso de inteligencia artificial generativa: articulación entre Roma I, Roma II y la gestión del riesgo

La cuestión de la responsabilidad jurídica derivada del uso de sistemas de IA generativa introduce un nivel adicional de complejidad en el análisis de estos contratos, en la medida en que obliga a articular dos planos normativos distintos, aunque estrechamente vinculados entre sí: el contractual, regido en principio por el Reglamento Roma I, y el extracontractual, sometido al Reglamento (CE) n.º 864/2007 (Roma II). Esta dualidad no constituye una novedad en el derecho internacional privado, pero adquiere una intensidad singular en el contexto de la IA, donde la producción del daño no siempre puede reconducirse con claridad a un simple incumplimiento contractual ni tampoco puede reducirse, sin más, a un supuesto clásico de responsabilidad extracontractual[21].

En efecto, los sistemas de IA generativa presentan una capacidad de producción de resultados que, aun siendo funcionalmente útil desde la perspectiva económica, puede generar perjuicios de naturaleza muy diversa. Entre ellos pueden mencionarse, por ejemplo, daños económicos derivados de información inexacta, errores en contextos especialmente sensibles, como el asesoramiento jurídico, médico o financiero, o la generación de contenidos discriminatorios. En todos estos supuestos, la delimitación entre responsabilidad contractual y extracontractual resulta particularmente compleja, sobre todo cuando el perjuicio afecta a sujetos que no son parte en el contrato de acceso al sistema.

Desde la perspectiva contractual, la responsabilidad del proveedor se encuentra, en la práctica, fuertemente condicionada por las cláusulas de limitación o exclusión de responsabilidad incluidas en las condiciones generales del servicio. Estas cláusulas suelen establecer que el sistema se proporciona «tal cual», sin garantías sobre la exactitud, fiabilidad o idoneidad de los resultados, y restringen la responsabilidad del proveedor a supuestos muy concretos. Sin embargo, la eficacia de este tipo de estipulaciones no puede analizarse de manera aislada, sino en conexión con las normas imperativas del ordenamiento aplicable, especialmente cuando el contrato presenta una vinculación significativa con el territorio de la Unión Europea.

En este sentido, las disposiciones del Derecho de la UE en materia de protección de consumidores, servicios digitales, protección de datos e IA pueden limitar o incluso neutralizar el efecto de determinadas cláusulas contractuales, en la medida en que estas contravengan obligaciones de carácter imperativo. La cuestión está lejos de ser puramente teórica. En un entorno en el que los principales proveedores de sistemas de IA operan a escala internacional, la pretensión de trasladar íntegramente el riesgo al usuario mediante cláusulas contractuales predispuestas entra en abierta tensión con la lógica protectora que inspira al ordenamiento europeo.

Ahora bien, más allá del plano contractual, la problemática de la responsabilidad se desplaza con frecuencia hacia el ámbito extracontractual. Cuando el daño se produce en la esfera de un tercero, la relación jurídica relevante deja de

21 Ortega Giménez, «El impacto extraterritorial…», pp. 221-239.

ser la que vincula al proveedor con el usuario y pasa a situarse en el terreno de las obligaciones extracontractuales. En tales casos, la determinación de la ley aplicable se rige, con carácter general, por el artículo 4.1 del Reglamento Roma II, conforme al cual la ley aplicable es la del país en el que se produce el daño, con independencia del país en el que se haya producido el hecho generador de dicho daño[22].

Este punto exige, no obstante, una precisión inmediata. No todos los daños derivados del uso de sistemas de inteligencia artificial generativa quedan comprendidos en el ámbito material del Reglamento Roma II. En particular, las obligaciones extracontractuales derivadas de violaciones de la intimidad o de los derechos de la personalidad, incluida la difamación, se encuentran expresamente excluidas por el artículo 1.2 g) del Reglamento. Por ello, resultaría técnicamente incorrecto tomar la difamación como ejemplo ordinario de aplicación de Roma II. Esta exclusión no resta complejidad al fenómeno; al contrario, la incrementa, pues pone de manifiesto que la fragmentación normativa derivada del uso de IA no solo se produce entre el plano contractual y el extracontractual, sino también dentro del propio ámbito de los daños extracontractuales.

Aun dejando al margen esos supuestos excluidos, la aplicación del artículo 4.1 Roma II plantea dificultades importantes en el contexto digital. La naturaleza transfronteriza de los servicios de IA generativa implica que el daño puede manifestarse de forma simultánea o sucesiva en una pluralidad de Estados, dificultando la localización de la *lex loci damni* en términos materialmente satisfactorios. Esta dificultad no es nueva en el contexto digital, pero se ve intensificada por la rapidez, escala y potencial automatización con la que los contenidos generados por sistemas de inteligencia artificial pueden circular y producir efectos en múltiples espacios jurídicos.

A ello se suma la cuestión, todavía más compleja, de la imputación del daño. A diferencia de los supuestos tradicionales, en los que el autor del hecho dañoso puede ser identificado con una relativa nitidez, en los sistemas de IA generativa la producción del resultado lesivo suele ser consecuencia de una interacción compleja entre distintos elementos, entre ellos el diseño y entrenamiento del modelo por parte del proveedor, la introducción de instrucciones o parámetros por el usuario y el propio funcionamiento probabilístico del sistema. Esta configuración dificulta la reconstrucción de una cadena causal lineal y obliga a desplazar el análisis hacia criterios más refinados de imputación del riesgo.

Desde una perspectiva jurídica, ello impide sostener, de forma general y abstracta, que la responsabilidad deba recaer de manera automática y exclusiva sobre uno solo de los sujetos implicados. La solución dependerá, necesariamente, de las circunstancias del caso concreto, del tipo de daño producido, del grado de intervención del usuario, del nivel de control efectivo del proveedor sobre el sistema y del marco normativo aplicable. Con todo, sí parece razonable afirmar que el proveedor no puede quedar, en principio, completamente al margen del

22 Domínguez Padilla, C., «La ley aplicable en la responsabilidad extracontractual derivada de sistemas de inteligencia artificial en el marco jurídico europeo», *Anuario Español de Derecho Internacional Privado*, n.º 25 (2025): 291-305.

análisis de responsabilidad, en la medida en que diseña, entrena, actualiza y pone a disposición el sistema que genera el riesgo[23]. Del mismo modo, el usuario puede incurrir en responsabilidad cuando utilice el sistema de forma negligente, contraria a Derecho o manifiestamente desviada de las condiciones de uso.

Esta posible concurrencia de responsabilidades no debe entenderse como una anomalía, sino como una consecuencia de la estructura misma del fenómeno. Los sistemas de IA generativa introducen una mediación tecnológica que dificulta la atribución del daño conforme a los modelos clásicos basados en una relación inmediata entre conducta y resultado. No se trata, sin embargo, de abandonar sin más los criterios tradicionales de responsabilidad, sino de reinterpretarlos a la luz de un entorno en el que la creación y distribución del riesgo aparecen tecnológicamente desagregadas.

En este contexto, el Reglamento Roma II desempeña una función esencial, al proporcionar el marco conflictual para la determinación de la ley aplicable a las obligaciones extracontractuales. No obstante, al igual que ocurre con el Reglamento Roma I, su aplicación no agota la complejidad del problema. La ley aplicable conforme al Reglamento debe coexistir, una vez más, con normas imperativas del Derecho de la Unión que pueden incidir decisivamente sobre la apreciación del riesgo, la configuración de los deberes de conducta o la delimitación misma de la responsabilidad, especialmente en ámbitos como la protección de datos personales, la no discriminación o la seguridad de los sistemas digitales.

Por otra parte, la creciente centralidad de la gestión del riesgo en la regulación europea de la IA introduce un elemento preventivo que transforma la propia lógica del análisis de responsabilidad. El Reglamento de Inteligencia Artificial no se limita a articular mecanismos de reacción frente al daño ya producido, sino que impone a determinados operadores una serie de deberes orientados a identificar, evaluar, documentar y mitigar los riesgos asociados al funcionamiento de los sistemas[24]. Esta orientación preventiva no sustituye a la responsabilidad, pero sí desplaza parcialmente el eje del sistema desde la reparación *ex post* hacia la evitación *ex ante* del daño, con implicaciones directas tanto en el plano material como en el conflictual[25].

En definitiva, la responsabilidad derivada del uso de sistemas de inteligencia artificial generativa no puede ser examinada desde una perspectiva exclusivamente contractual ni exclusivamente extracontractual. Se trata de un ámbito en el que confluyen ambos planos, generando una interacción compleja entre el Reglamento Roma I y el Reglamento Roma II, a la que se superpone además la incidencia de normas de aplicación necesaria del Derecho de la Unión. Esta configuración pone de manifiesto, una vez más, que los instrumentos clásicos del derecho internacional privado siguen siendo operativos, pero ya no resultan

23 Merchán Murillo, «Ley aplicable en la contratación automatizada», pp. 71-94.

24 Ortega Giménez, A., «La responsabilidad civil extracontractual por un uso inadecuado de sistemas de inteligencia artificial y el necesario carácter complementario entre el Reglamento europeo de Inteligencia Artificial y el Reglamento General de Protección de Datos», *Unión Europea Aranzadi*, n.º 1 (2026).

25 Domínguez Padilla, «La ley aplicable…», pp. 291-305.

suficientes por sí solos para ordenar de manera completa este tipo de relaciones. De ahí la necesidad de una aproximación integrada que atienda no solo a la ley aplicable en sentido estricto, sino también a la estructura del riesgo, a la pluralidad de sujetos intervinientes y a la densidad normativa del ecosistema digital contemporáneo.

VI. Conclusiones

Los contratos de uso de IA generativa ponen en crisis la suficiencia de las categorías contractuales clásicas, al no articularse en torno a una prestación cerrada, sino al acceso a una capacidad generativa cuya ejecución efectiva permanece, en gran medida, fuera del control del usuario.

Esta singularidad repercute directamente en la operación de calificación jurídica. Aunque, en principio, tales contratos puedan reconducirse a la categoría de prestación de servicios del artículo 4.1 b) del Reglamento Roma I, dicha solución resulta más operativa que plenamente satisfactoria desde un punto de vista dogmático.

La aplicación de la ley del país de residencia habitual del prestador revela límites evidentes en este ámbito, pues existe con frecuencia una profunda disociación entre la localización formal del proveedor y el lugar en que se concentran los efectos jurídicos y económicos del contrato.

Sin embargo, la respuesta a esta insuficiencia no reside tanto en una utilización expansiva de la cláusula de escape del artículo 4.3 del Reglamento Roma I, cuanto en la creciente incidencia de normas de aplicación necesaria del Derecho de la UE, que condicionan estructuralmente la relación contractual con independencia de la ley designada.

Reglamentos como el RGPD, el Reglamento de Inteligencia Artificial, el Reglamento de Servicios Digitales o el Reglamento de Datos muestran que, en los contratos de uso de inteligencia artificial generativa, la *lex contractus* sigue siendo relevante, pero deja de constituir el único eje normativo de la relación.

La complejidad del fenómeno se intensifica en el plano de la responsabilidad, donde la articulación entre Roma I y Roma II se vuelve ineludible. Los daños derivados del uso de estos sistemas no siempre pueden reconducirse con claridad a un incumplimiento contractual, ni tampoco quedan íntegramente absorbidos por los modelos clásicos de responsabilidad extracontractual.

En particular, la estructura tecnológicamente mediada de estos sistemas dificulta una imputación lineal del daño y obliga a atender a la pluralidad de sujetos intervinientes, a la contribución de cada uno a la creación del riesgo y a la creciente centralidad de los deberes preventivos de identificación, evaluación y mitigación.

En definitiva, los contratos de uso de inteligencia artificial generativa no invalidan los instrumentos clásicos del derecho internacional privado, pero sí ponen de manifiesto sus límites funcionales y exigen una reinterpretación

integrada de sus categorías a la luz de la densidad normativa y de la complejidad técnica propias del ecosistema digital contemporáneo.

Referencias bibliográficas

Calvo Caravaca, A. L. y Carrascosa González, J., *Tratado crítico de Derecho Internacional Privado. Vol. VI. Derecho de los Negocios Internacionales II*. Edisofer Libros Jurídicos, 2024.

Calvo Caravaca, A. L., «El Reglamento Roma I Sobre la Ley Aplicable a las Obligaciones Contractuales: Cuestiones Escogidas». *Cuadernos de derecho transnacional*, vol. 1, n.º 2 (2009).

Castelló Pastor, J. J., Guerrero Pérez, A. y Martínez Pérez, M. *Derecho de la contratación electrónica y comercio electrónico en la Unión Europea y en España*. Tirant lo Blanch, 2021.

De Miguel Asensio, P. A. «Cláusulas de elección del derecho aplicable». *Cláusulas en los contratos internacionales: redacción y análisis*, dir. S. Sánchez Lorenzo, 2021.

De Miguel Asensio, P. A. «Contratación comercial internacional». *Derecho de los negocios internacionales*, coord. J. C. Fernández Rozas, R. Arenas García, P. A. de Miguel Asensio, 2024.

De Miguel Asensio, P. A. «Reglamento (UE) 2024/1689 de Inteligencia Artificial y Derecho Internacional Privado». *Revista española de derecho internacional*, vol. 77, n.º 1 (2025).

De Miguel Asensio, P. A. «Reglamento General de Protección de Datos: Coordinación entre la aplicación pública y la aplicación privada». *La Ley Unión Europea*, n.º 112 (2023).

Deskoski, T. y Dokovski, V. «Lex Contractus for Specific Contracts under the Rome I Regulation». *Iustinianus Primus L. Rev.*, vol. 10 (2019).

Domínguez Padilla, C. «La ley aplicable en la responsabilidad extracontractual derivada de sistemas de inteligencia artificial en el marco jurídico europeo». *Anuario Español de Derecho Internacional Privado*, n.º 25 (2025): 291-305.

Domínguez Padilla, C. «Leyes de policía en la contratación con sistemas de inteligencia artificial», *Anuario Español de Derecho Internacional Privado*, n.º 24 (2024): 107-129.

Jacquet, J. M., «La aplicación de las leyes de policía en materia de contratos internacionales». *Anuario español de Derecho internacional privado*, vol. 10 (2010): 35-48.

Merchán Murillo, A. «Ley aplicable en caso de responsabilidad a las plataformas digitales». *Plataformas digitales: aspectos jurídicos*, dir. A. Martínez Nadal, 2021.

Merchán Murillo, A., «Ley aplicable en la contratación automatizada». *Derecho internacional privado en acción: algunas cuestiones: IV Foro europeo de Derecho internacional privado*, coord. A. Fernández Pérez, 2024.

Merchán Murillo, A. *Derecho internacional privado y nuevas tecnologías*. Colex, 2026.

Muñoz Vela, J. M. «Inteligencia artificial generativa. Desafíos para la propiedad intelectual». *Revista de Derecho de la UNED*, n.º 33 (2024).

Ortega Giménez, A. «Artificial intelligence, smart contracts and private international law in Spain». *Los retos que plantea la inteligencia artificial a la administración pública*, coord. L. S. Heredia Sánchez, 2025.

Ortega Giménez, A. «El impacto extraterritorial del Reglamento europeo de inteligencia artificial», *Anuario de la Facultad de Derecho*, n.º 17 (2024): 221-239.

Ortega Giménez, A. «Evolución tecnológica, protección de datos de carácter personal y derecho internacional privado», *Derecho internacional privado en acción: algunas cuestiones: IV Foro europeo de Derecho internacional privado*, coord. A. Fernández Pérez, 2024.

Ortega Giménez, A. «La responsabilidad civil extracontractual por un uso inadecuado de sistemas de inteligencia artificial y el necesario carácter complementario entre el Reglamento europeo de Inteligencia Artificial y el Reglamento General de Protección de Datos». *Unión Europea Aranzadi*, n.º 1 (2026).

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 108-120

A BUSCA PELA ADEQUAÇÃO: TRANSFERÊNCIA INTERNACIONAL DE DADOS UE-BRASIL PÓS JURISPRUDÊNCIA SCHREMS

THE SEARCH FOR ADEQUACY: INTERNATIONAL DATA TRANSFER EU-BRAZIL POST-SCHREMS JURISPRUDENCE

Francisco Pedro Gonçalves Carvalho de Sousa
Universidade do Minho

Resumo

O presente estudo objetiva analisar a aplicação extraterritorial do Regulamento Geral de Proteção de Dados no contexto da transferência internacional, bem como de qual modo se deu a influência dos acórdãos *Schrems* I e II na instituição dos parâmetros reguladores desse fluxo para territórios de fora da UE. Adiante, é sopesado qual o grau de influência desses julgamentos e do RGPD na feitura da Lei Geral de Proteção de Dados brasileira, ao se examinar os fluxos de dados pessoais nas transferências internacionais e a aplicabilidade dos mecanismos jurídicos que os regulam. Por fim, busca-se dilucidar quanto à possibilidade de reconhecimento da sintonia do arcabouço jurídico brasileiro com os pontos essenciais reclamados pela União Europeia na identificação do nível adequado de proteção de dados, a fim de aquiescer a transmissão juridicamente segura de dados pessoais entre as partes.

Palavras-chave

Transferência internacional de dados pessoais, Regulamento Geral de Proteção de Dados, Lei Geral de Proteção de Dados, conjuntura do fluxo UE-Brasil

Abstract

The present study aims to analyze the extraterritorial application of the General Data Protection Regulation in the context of international transfer, as well as how the influence of the Schrems I and II judgments on the establishment of the parameters regulating this flow to territories outside the EU occurred. Furthermore, it is considered the degree of influence of these judgments and the GDPR in the making of the Brazilian General Data Protection Law, when examining the flows of personal data in international transfers and the applicability of the legal mechanisms that regulate them. Finally, it is sought to clarify the possibility of recognition of the harmony of the Brazilian legal framework with the essential points demanded by the European Union in identifying the appropriate level of data protection, in order to consent to the legally secure transmission of personal data between the parties.

Keywords

International transfer of personal data, General Data Protection Regulation, Brazilian General Data Protection Law, EU-Brazil flow conjuncture

I. Introdução

Na atual sociedade hiperconectada, os dados ocupam lugar de destaque na economia da informação, uma vez que esta passou a ocupar um papel central na economia digital contemporânea, ao funcionar tanto como bem comercializável quanto como elemento essencial nos processos produtivos (Rodrigues, 2016). Em modelos econômicos informacionais, os dados assumem uma importância basilar, atuando como as unidades fundamentais de valor.

Sendo assim, no contexto diplomático entre Brasil e União Europeia, os mecanismos e critérios de proteção de dados tornam-se fundamentais para viabilizar a aplicação extraterritorial das normas brasileiras e europeias. Esses instrumentos visam a estabelecer um padrão mínimo global de proteção, incentivando os parceiros econômicos a se adaptarem às exigências de proteção de dados, para enfim assegurar o fluxo internacional de informações.

Inicialmente, o presente artigo aborda a extraterritorialidade adotada pelo Regulamento Geral de Proteção de Dados, bem como os critérios adotados para efetivar a transferência internacional de dados. No segundo tópico, de posse da jurisprudência *Schrems*, são perquiridos os impactos da decisão nas transferências internacionais de dados pessoais entre a União Europeia e outros países. Por sua vez, o terceiro tópico investiga se o Brasil pode se colocar como apto a receber a decisão de adequação aos padrões do RGPD, a fim de assegurar um fluxo livre de dados com a UE.

II. Extraterritorialidade e transferência internacional de dados no RGPD

No âmbito do direito internacional, reconhece-se que cada Estado soberano possui o direito de acessar os dados que se encontram sob sua jurisdição para cumprir suas finalidades regulatórias, especialmente no que se refere ao combate a crimes, à gestão de riscos e à proteção da segurança nacional. Em geral, os países utilizam todas as informações disponíveis dentro de seus limites territoriais ou vinculadas aos seus cidadãos. A atuação além das fronteiras é mais restrita a áreas específicas, como ocorre nas políticas de concorrência aplicadas por Estados Unidos e União Europeia (Schweighofer, 2017).

Dada a internacionalização da troca de dados, soluções mais maleáveis são imprescindíveis para que uma abordagem mais segura dos princípios do direito internacional seja feita. Para Schweighofer (2017), liberdade de expressão, soberania territorial, liberdade de comunicação, direito à privacidade e neutralidade da internet são alguns exemplos desses princípios basilares.

O RGPD dá ênfase e escancara a possibilidade de sua aplicação extraterritorial ao mencionar em seu artigo 3º a possibilidade de aplicação ao tratamento de dados pessoais quando o tratamento ocorrer no âmbito das operações de uma entidade controladora ou processadora situada na União Europeia, ainda que o processamento propriamente dito aconteça fora do território europeu, bem como quando houver tratamento de dados de pessoas localizadas na União por parte de controladores ou processadores estabelecidos fora dela, desde que essa atividade esteja relacionada à oferta de bens ou serviços para esses indivíduos ou ao acompanhamento de seu comportamento enquanto se encontram no território da União.

Ainda, o número 3 do mesmo artigo esclarece que «o presente regulamento aplica-se ao tratamento de dados pessoais por um responsável pelo tratamento estabelecido não na União, mas num lugar em que se aplique o direito de um Estado-Membro por força do direito internacional público».

Percebe-se assim a intenção do legislador europeu a conferir maior capacidade na proteção de dados pessoais, ao ampliar o escopo territorial de aplicação do Regulamento.

Relativiza-se, assim, a localização geográfica dos utilizadores dos dados, ao passo que a jurisdição aplicável é ampliada de forma considerável. Para Kuner (2019), essa ampliação é aplicada apenas quando existe algum vínculo com a União Europeia, seja pela oferta de produtos ou serviços ao seu público, seja em razão das atividades desenvolvidas que mantenham relação com o bloco.

Desta forma, a transferência internacional de dados está intrinsicamente ligada à extraterritorialidade das normas de proteção de dados. Nesse contexto, Cravo (2024, p. 196) aponta que a União Europeia:

> (…) por meio de seu robusto arcabouço jurídico, estabelece garantias de uma proteção adequada de dados durante essas transferências, funcionando como uma forma de exportação do seu Direito. Isso significa que, ao exigir que países terceiros assegurem níveis adequados de proteção de dados, a legislação da UE não apenas regula o tratamento de dados dentro de suas fronteiras, mas também exerce influência sobre a legislação de outras jurisdições.

Ainda, de acordo com o exposto por Svantesson (2011, como citado em Filho & Uehara, 2019), a regulação das transferências internacionais de dados deve adotar uma postura neutra do ponto de vista tecnológico. Isso implica reconhecer que os fluxos de dados só podem ocorrer conforme determinados requisitos destinados a garantir a proteção dos direitos dos titulares sejam atendidos. Além disso, para preservar o nível de salvaguarda, é necessário assegurar que o responsável pelo envio dos dados continue sujeito a deveres claros e que sejam conduzidas análises periódicas de risco, capazes de verificar se a transferência é capaz de gerar prejuízo ao indivíduo envolvido.

De tal sorte que o RGPD adota como primado basilar o condicionamento da transferência de dados pessoais para tratamento em país terceiro, ou por organização internacional, se forem respeitadas as exigências previstas no próprio regulamento. O objetivo é garantir que as pessoas físicas mantenham um nível adequado de proteção, inclusive nas eventuais transferências subsequentes que possam ocorrer a partir do destinatário inicial para outros países ou organizações internacionais (RGPD, 2016).

Assim, o RGPD mantém três modelos centrais para disciplinar as transferências internacionais de dados, a saber, o reconhecimento de que o país de destino oferece um nível de proteção considerado adequado (art. 45º), a adoção de garantias suficientes para assegurar essa proteção (art. 46º) e as derrogações de caráter específico (art. 49º) (Ramalho, 2020).

III. A jurisprudência Schrems: caso paradigmático e o nível de proteção adequado

1. Schrems I

É importante enfatizar a relevância do Tribunal de Justiça da União Europeia - TJUE na interpretação do que se pode entender como nível adequado de proteção, conforme positivado no artigo 45º do RGPD. A mencionada contribuição restou evidenciada no caso *Schrems* I (C-362/14), em que o tribunal analisou se os Estados Unidos ofereciam um padrão de proteção compatível com o sistema europeu de salvaguarda de dados pessoais (Domínguez, 2019).

In casu, após a divulgação por parte de Edward Snowden de que os Estados Unidos, por meio da Agência de Segurança Nacional (NSA), empregava massiva atividade de espionagem contra seus cidadãos, o austríaco *Maximillian Schrems* argumentou que a transferência de dados do Facebook para os Estados Unidos era ilegal, pois essas transferências haviam sido interceptadas pela NSA. Com isso, em 2013, *Schrems* apresentou uma reclamação à autoridade irlandesa responsável pela fiscalização da proteção de dados, solicitando que fosse impedida a transferência de suas informações pessoais para os Estados Unidos. O caso restou conhecido como *Schrems* v. *Irish Data Protection Commissioner*.

Após o caso chegar ao TJUE, a Corte tomou como base, sobretudo, os artigos 25 e 28 da Diretiva 95/46/CE (antigo *codex* que regulamentava o processamento de dados na UE em tempos pré-RGPD), além dos artigos 7, 8 e 47 da Carta dos Direitos Fundamentais da União Europeia. O artigo 25 da Diretiva previa que o envio de dados a países fora da UE só poderia ocorrer se o destino oferecesse garantias suficientes de proteção e se a Comissão tivesse competência para emitir uma decisão reconhecendo essa adequação. O artigo 28, por sua vez, determinava que cada Estado-membro instituísse ao menos uma autoridade pública encarregada de supervisionar o cumprimento das normas (Osman & Soares, 2020).

O acórdão exarado pelo Tribunal é centrado em três temas-chave: 1) até que ponto as autoridades nacionais de proteção de dados podem atuar quando existe uma decisão de adequação emitida pela Comissão Europeia; 2) como deve ser compreendido o que constitui um nível adequado de proteção e 3) se a Decisão 2000/520, responsável por declarar que os Estados Unidos possuíam nível de proteção adequado, permanecia válida diante das questões levantadas (Araújo, 2017).

De tal sorte que Araújo (2017) esclarece que a resolução de cada um dos temas centrais do julgamento se deu da seguinte forma: 1) O Tribunal afirma que o artigo 28 da antiga Diretiva de Proteção de Dados impõe às autoridades nacionais o dever contínuo de avaliar se uma transferência internacional realizada a partir do Estado-membro está em conformidade com as exigências previstas na norma; 2) decidiu-se que a antiga Diretiva de Proteção de Dados deve ser compreendida de modo a garantir uma tutela plena e eficaz dos direitos e liberdades previstos na Carta, de modo a assegurar um patamar elevado de proteção às pessoas. Para a Corte, a avaliação sobre se um país terceiro oferece um nível de proteção essencialmente equivalente ao da União Europeia deve levar em conta

a realidade prática; 3) o Tribunal concluiu que a Decisão 2000/520 não garantia um nível de proteção compatível com o exigido e contrariava as condições previstas no artigo 25º, n.º 6, da Diretiva 95/46/CE. Por esse motivo, tal decisão deveria ser considerada inválida.

A partir do acórdão *Schrems*, a transferência de dados EU-EUA restou tolhida, e o fluxo de dados entre os territórios ficou restringido às outras possibilidades previstas pela legislação, como o uso de cláusulas contratuais, de normas corporativas vinculantes ou das exceções especiais estabelecidas no artigo 26 da Diretiva 95/46/CE (Araújo, 2017). Com isso, em fevereiro de 2016, depois de um entendimento firmado entre a União Europeia e os Estados Unidos (*EU-US Privacy Shield*), a Comissão Europeia emitiu uma nova decisão de adequação.

Essa decisão considerava que as exigências apontadas pelo TJUE no caso *Schrems*, especialmente no que diz respeito às salvaguardas necessárias para as transferências internacionais de dados, estavam atendidas, garantindo assim o nível de proteção requerido pelo RGPD para o envio de informações pessoais da UE aos EUA (Osman & Soares, 2020). Portanto, para Araújo (2017), além de elevar o grau de proteção conferido aos dados europeus tratados em países fora do bloco, preservou-se algum espaço para que estas nações ajustem a aplicação das regras conforme suas próprias tradições e culturas jurídicas.

Ademais, a leitura conjunta dos artigos 44º e 45º do RGPD é esclarecedora, e a partir do caso *Schrems*, podemos depreender que o princípio norteador da transferência internacional de dados é a exigência de que o país terceiro garanta o mesmo nível de proteção de dados assegurado pelo RGPD. O próprio RGPD, em seu Considerando (104), é cristalina ao integrar a interpretação dada pelo TJUE de que o princípio de proteção adequado é efetivado quando o outro país puder «assegurar um nível adequado de proteção essencialmente equivalente ao assegurado na União» (RGPD, 2016).

2. Schrems II

Após o *Privacy Shield*, as empresas baseadas nos Estados Unidos puderam aderir aos princípios e termos do acordo a fim de executar a transferência internacional de dados de cidadãos localizados na União Europeia para o país americano. No entanto, em 16 de julho de 2020, o Tribunal de Justiça da União Europeia, ao julgar o processo C-311/18 (*Schrems* II), entendeu que o arcabouço jurídico americano vai de encontro ao RGPD e revogou a decisão de adequação que fundamentava o *Privacy Shield*.

Conforme Moreira (2023), o Tribunal concluiu que as limitações impostas pela legislação americana ao acesso e uso de dados pessoais por entidades públicas (para fins de segurança nacional), não atendiam aos requisitos de necessidade e proporcionalidade exigidos pela legislação europeia. Ainda, verificou-se que a lei americana não oferecia garantias legais adequadas para que os indivíduos afetados pudessem buscar recurso em face das ingerências com seus direitos, perante uma entidade com preceitos equivalentes às estipuladas no Artigo 47º da Carta dos Direitos Fundamentais da União Europeia.

Desta feita, no pertinente ao uso de cláusulas contratuais padrão (positivadas no art. 46º do RGPD), o TJUE reconheceu que, previamente à transferência de dados pessoais para países terceiros, o controlador tem a obrigação de realizar uma análise adicional acerca das garantias de proteção previstas no RGPD. Caso esta análise não tome forma, as autoridades de proteção de dados dos países membros da UE poderão barrar a transferência de dados para países como os Estados Unidos (Osman & Soares, 2020).

3. Impactos práticos do Schrems II

A decisão *Schrems* II alterou profundamente o paradigma das transferências internacionais de dados ao expor as insuficiências dos mecanismos que buscavam equiparar a proteção externa ao padrão do RGPD. Ao invalidar o *Privacy Shield*, o TJUE não apenas removeu um pilar fundamental do fluxo transfronteiriço, mas instaurou um cenário de insegurança jurídica para organizações globais.

O acórdão transferiu aos agentes privados um encargo técnico-jurídico desproporcional: a obrigação de avaliar, em cada operação, se o ordenamento do país de destino oferece garantias equivalentes às europeias. Essa exigência é particularmente problemática na prática, dada a opacidade dos sistemas de vigilância e a atuação de serviços de inteligência estrangeiros, elementos que fogem ao controle e à capacidade de análise das empresas.

Para viabilizar as transferências, o Tribunal passou a exigir medidas suplementares, como criptografia robusta e pseudonimização. No entanto, questiona-se a eficácia dessas salvaguardas, argumentando que soluções tecnológicas, por mais avançadas que sejam, raramente conseguem neutralizar integralmente ingerências estatais desproporcionais em determinados contextos jurisdicionais.

Embora essencial para a salvaguarda dos direitos fundamentais, o impacto operacional da decisão é contencioso. Isso pois o rigor do *Schrems II* cria barreiras ao comércio digital e desincentiva a inovação em uma economia dependente da circulação de dados. Em última análise, o acórdão evidencia que o modelo atual de conformidade caso a caso é insuficiente, reforçando a urgência de soluções multilaterais que harmonizem a privacidade com as dinâmicas da economia digital contemporânea.

IV. O nível adequado de proteção de dados pelo Brasil

1. A LGPD e a regulação brasileira de transferência de dados

A Lei Geral de Proteção de Dados – LGPD é a legislação brasileira dedicada a regular a protecção e tratamento de dados pessoais. O capítulo V, que engloba os artigos 33 a 36, regula a transferência internacional de dados. Conforme exposto por Ramalho (2020), as similaridades com o RGPD são patentes, e a LGPD adota em seu artigo 33 três modalidades principais de proteção para as transferências internacionais de dados: a possibilidade de reconhecer que o país de destino oferece um nível adequado de proteção (inciso I); a utilização de mecanismos

que assegurem o cumprimento das normas da LGPD (inciso II); e a aplicação de exceções específicas previstas na própria lei (incisos III a IX). Cabe à Autoridade Nacional de Proteção de Dados - ANPD avaliar o nível de proteção do lugar de destino.

Em sua decisão, a ANPD deve verificar se as normas adotadas pelos países destinatários estão em sintonia com a legislação brasileira. Segundo Osman & Soares (2020), caso essa compatibilidade não exista, os dados pessoais provenientes do Brasil só poderão ser enviados ao exterior mediante a validação da ANPD, por meio de instrumentos alternativos, como cláusulas contratuais específicas, consentimento claro e informado do titular, códigos de conduta, selos e mecanismos de certificação.

No entanto, Carvalho (2019) pondera que o legislador brasileiro parece atribuir mais importância à observância dos princípios da LGPD por parte de países estrangeiros do que à exigência de existência de legislação específica e unificada de proteção de dados. Assim, o conjunto normativo aplicável tenciona a se manifestar de maneira dispersa dentro do ordenamento jurídico do país destinatário.

Ao passo que a LGPD estabelece que em caso de uso de instrumentos destinados à transferência de dados, como as cláusulas-padrão, cabe ao controlador assegurar que suas orientações e as regras aplicáveis à proteção dos dados do titular sejam devidamente seguidas, devendo ainda comunicar ao operador as instruções para a realização do tratamento adequado (Franco, 2020).

Acerca dessa temática, a Resolução CD/ANPD n.º 19, de 23 de agosto de 2024, da ANPD, define as regras para a transferência internacional de dados e fixa o conteúdo das cláusulas-padrão contratuais, ao oferecer diretrizes objetivas para esse tipo de operação. O objetivo principal é assegurar que o envio de dados pessoais ao exterior seja realizado com segurança e em conformidade com as exigências de proteção previstas na LGPD.

A Resolução estabelece que a transferência internacional de dados deve seguir a LGPD, e exige, entre outras diretrizes, a garantia de um nível de proteção equivalente ao nacional no país de destino, a adoção de procedimentos simples e compatíveis com boas práticas internacionais, a promoção do livre fluxo transfronteiriço com confiança e desenvolvimento, a responsabilização e prestação de contas pela observância dos direitos dos titulares, a implementação de medidas de transparência sobre a transferência, e a adoção de boas práticas e medidas de segurança apropriadas à natureza dos dados e riscos envolvidos.

Com base nesse conteúdo, é perceptível que os fundamentos do *Schrems* II foram incorporados pela LGPD, de forma a instituir os padrões e regramentos definidores da validade da fluência de dados para além do território brasileiro. É dizer, assim como ocorreu no caso paradigmático aqui analisado, a legislação atribui ao controlador uma responsabilidade significativa, que deve ser rigorosamente cumprida e pode ser objeto de avaliação tanto por autoridades administrativas quanto pelo Poder Judiciário.

2. A adequação brasileira

Para Viola (2019, p. 8), eventual reconhecimento pela União Europeia acerca da proteção dada pela LGPD ao Brasil seria benéfico pois

> permitirá que haja o livre fluxo de dados com a União Europeia, com potencial impacto econômico positivo, já que a economia relacionada ao mercado de dados deverá representar 5,4% do PIB da União Europeia até o ano de 2025.

Pondera Ramalho (2020) que, conforme analisamos os critérios fundamentais previstos no artigo 45, números 1, 2 e 3, do RGPD, a saber, o respeito ao Estado de Direito, a atuação efetiva e independente de uma autoridade de controle e os compromissos internacionais assumidos pelo país, o estado atual do sistema brasileiro vai ao encontro do ordenamento europeu. Sobre isso, Ramalho (2020, p. 105) aduz:

> Sendo o Brasil um país democrático, com uma ordem jurídica fundamentada na Dignidade da Pessoa Humana e sendo signatário das principais convenções internacionais sobre direitos humanos tais como: a Declaração Universal dos Direitos Humanos (1948), a Declaração Americana de Direitos e Deveres do Homem (1948), o Pacto Internacional de Direitos Civis e Políticos (1966) e a Convenção Americana de Direitos Humanos, ou Pacto de São José da Costa Rica (1969), não deve haver dúvida razoável quanto ao primeiro critério.
>
> Ademais, o Brasil já inaugurou seu pleito para se filiar à OCDE, sendo hoje o país que cumpre o maior número de requisitos para entrada no organismo internacional. Já no âmbito internacional da privacidade e da proteção de dados, a aprovação da LGPD concedeu ao País o status de membro observador da Convenção 108/1981.

Por fim, a Lei n.º 14.460, de 25 de outubro de 2022, transformou a ANPD em entidade independente, de forma a cumprir mais um requisito posto pelo RGPD. À luz destes acontecimentos, em 5 de setembro de 2025, a Comissão Europeia divulgou a versão preliminar da futura decisão de adequação, com fins de reconhecer que o Brasil assegura nível de proteção de dados pessoais equivalente ao balizado no RGPD (ANPD, 2025). O relatório é enfático ao afirmar:

> Para efeitos do Artigo 45 do Regulamento (UE) 2016/679, o Brasil assegura um nível adequado de proteção para os dados pessoais transferidos da União Europeia para responsáveis pelo tratamento e operadores de dados no Brasil, sujeitos à Lei Geral de Proteção de Dados (LGPD) (Comissão Europeia, 2025, p. 52).

Somado a isso, ANPD atualmente encontra-se a trabalhar na sua emissão de decisão de adequação, de forma a possibilitar o reconhecimento da equivalência da legislação europeia com os preceitos da LGPD. Outrossim, conforme relata a ANPD (2025), a UE iniciará os trâmites finais para conceder o status de adequação ao Brasil. Ao final, o país passará a integrar o grupo de 16 países

já detentores de decisão de adequação, tais como, Reino Unido, Canadá, Japão, Coreia do Sul, Argentina e Uruguai.

Certo é que a LGPD guarda uma dívida teórica evidente com o modelo europeu, especialmente no regime das transferências internacionais. Ambas as normas convergem ao eleger a equivalência do nível de proteção no país de destino como critério central, estruturando-se sobre um sistema de salvaguardas que inclui cláusulas contratuais padrão e normas corporativas vinculantes.

Contudo, a identidade estrutural não apaga divergências substantivas. O RGPD caracteriza-se por uma densidade normativa superior, fruto da maturidade institucional do bloco europeu, enquanto a legislação brasileira adota uma postura mais flexível. Essa diferença traduz-se em uma maior margem de discricionariedade para a autoridade brasileira na avaliação da conformidade de terceiros Estados e dos instrumentos de transferência utilizados.

O arranjo das autoridades de controle também revela estágios distintos de desenvolvimento. Se o ecossistema europeu se beneficia da atuação coordenada entre autoridades nacionais, o modelo brasileiro ainda atravessa um processo de consolidação. A questão central para uma futura decisão de adequação pela União Europeia reside na efetividade prática do sistema brasileiro. Pairam dúvidas sobre se a cultura de proteção de dados e a capacidade sancionatória da ANPD alcançam, de fato, a equivalência exigida por Bruxelas. Portanto, embora o alinhamento formal entre LGPD e RGPD seja robusto, a chancela europeia dependerá menos do texto legal e mais da robustez da fiscalização e da aplicação concreta da norma no ambiente empresarial brasileiro.

V. Conclusão

A fim de asseverar a privacidade de dados dos cidadãos e assegurar as garantias previstas no RGPD, a União Europeia faz uso da aplicação da extraterritorialidade. Sob esse mote, relativiza-se a localização geográfica do sujeito e o campo de aplicação do Regramento é ampliado, desde que certos requisitos sejam cumpridos e haja vínculo com a União. Entre esses requisitos, merecem maior destaque as decisões de adequação emitidas pela Comissão Europeia.

Os casos *Schrems* I e II são notáveis, dado seu papel em elucidar que o princípio do nível adequado de proteção obriga o país terceiro a garantir, de forma concreta, um grau de tutela das liberdades e direitos fundamentais que seja essencialmente equivalente ao existente na União Europeia. Para mais, o TJUE também reconheceu a responsabilidade do controlador em adotar medidas adicionais sempre que necessário, a fim de assegurar a conformidade da proteção dos dados transferidos com o RGPD.

No mais, a LGPD foi verdadeiro marco no contexto brasileiro de proteção de dados. Sua forte inspiração no RGPD se dá pela integração de mecanismos semelhantes para resguardar informações pessoais em transferências internacionais. A lei permite e egula a transferência de dados transfronteiriça, bem como adota fundamentos alinhados àqueles destacados pelo TJUE no *Schrems* II.

Quanto às transferências UE-Brasil, após o Brasil atingir os requisitos previstos, a UE encontra-se prestes a emitir a decisão de adequação definitiva, e o Brasil, de forma recíproca, planeja reconhecer a equivalência da legislação europeia com os mandamentos da LGPD.

No entanto, a eficácia de um regime de proteção de dados transcende a mera conformidade normativa; sua consolidação exige que a legislação formal se traduza em práxis institucional. Mais do que a letra da lei, a segurança jurídica depende da internalização desses princípios pelos agentes econômicos e de uma aplicação administrativa que assegure a efetividade dos direitos previstos.

É provável que novos questionamentos jurisdicionais continuem a tensionar e redefinir os parâmetros de conformidade globais. Em última análise, a viabilidade de um fluxo de dados seguro e perene não será alcançada por esforços unilaterais isolados, mas pela convergência progressiva dos regimes jurídicos e pelo fortalecimento da cooperação internacional, elementos indispensáveis para harmonizar a privacidade com as exigências da economia digital.

E se é assim, é forçoso reconhecer que a maneira mais eficaz de garantir um fluxo de dados seguro envolve a adoção de soluções tecnológicas e práticas de privacidade incorporadas desde o desenho dos sistemas, além da celebração de acordos bilaterais estratégicos e à construção de uma matriz global de proteção de dados.

Referências bibliográficas

Araújo, A. M. R. (2017). As transferências transatlânticas de dados pessoais: o nível de proteção adequado depois de Schrems. *Revista Direitos Humanos e Democracia*, *5*(9), 201-236. https://doi.org/10.21527/2317-5389.2017.9.201-236

Autoridade Nacional de Proteção de Dados (ANPD). (2024, 23 de agosto). Resolução CD/Anpd nº 19. Aprova o Regulamento de Transferência Internacional de Dados e o conteúdo das cláusulas-padrão contratuais. *Diário Oficial da União*, Seção 1, pp. 123-237. https://www.in.gov.br/en/web/dou/-/resolucao-cd/anpd-n-19-de-23-de-agosto-de-2024-580095396

Autoridade Nacional de Proteção de Dados (ANPD). (2025, 5 de setembro). União Europeia divulga versão preliminar de decisão de adequação. *Portal Gov.br*. https://www.gov.br/anpd/pt-br/assuntos/noticias/uniao-europeia-divulga-versao-preliminar-de-decisao-de-adequacao

Brasil. (2018). Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). *Presidência da República, Casa Civil*. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

Brasil. (2022). Lei nº 14.460, de 25 de outubro de 2022. Transforma a Autoridade Nacional de Proteção de Dados (ANPD) em autarquia de natureza especial; altera as Leis nºs 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais), e 13.844, de 18 de junho de 2019; e revoga dispositivos da Lei nº 13.853, de 8 de julho de 2019. *Presidência da República, Casa Civil*.

https://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2022/Lei/L14460.htm

Carvalho, A. G. P. de. (2019). Transferência Internacional de dados na Lei Geral de Proteção de Dados — Força normativa e efetividade diante do cenário transnacional. Em G. tepedino, A. Frazão, M. Oliva (Coord.), Lei Geral de Proteção de Dados Pessoais e suas repercussões no Direito Brasileiro. 1ª edição, Thompson Reuters Brasil, São Paulo, 2019. p. 621-646.

Comissão Europeia. (2000). *Decisão 2000/520/CE da Comissão, de 26 de julho de 2000, nos termos da Directiva 95/46/CE do Parlamento Europeu e do Conselho e relativa ao nível de protecção assegurado pelos princípios de "porto seguro" e pelas respectivas questões mais frequentes (Faq) emitidos pelo Department of Commerce dos Estados Unidos da América*. https://eur-lex.europa.eu/legal-content/PT/Txt/Pdf/?uri=Celex:32000D0520

Comissão Europeia. (2016). *EU Commission and United States agree on new framework for transatlantic data flows: EU-US Privacy Shield* (Comunicado de Imprensa IP/16/216). https://europa.eu/rapid/press-release_IP-16-216_en.htm

Comissão Europeia. (2025). *Draft Implementing Decision pursuant to Regulation (EU) 2016/679 of the European Parliament and of the Council on the adequate level of protection of personal data in Brazil*. Comissão Europeia. https://commission.europa.eu/document/download/f5aee532-70bf-41b1-a94a-8e294a528f6a_en?filename=Draft%20Adequacy%20Decision%20-%20Brazil%20%20LGPD%20-%20FINAL%20-%20September%202025.pdf

Cravo, D. C. (2024). Proteção de dados pessoais na era digital: transferência internacional e efeitos extraterritoriais. Em A. R. Pilati & M. R. L. Aguirre (Eds.), *Inovação e Sustentabilidade no Direito III* (pp. 183-205). Free Press.

Domínguez, C. (2020). *La noción de «consumidor» en internet: El asunto C-498/16, Maximilian Schrems y Facebook Ireland Limited*. Universidad Carlos III de Madrid.

Filho, P. C. T., & Uehara, L. F. (2020). *Transferência internacional de dados pessoais: uma análise crítica entre o regulamento geral de proteção de dados pessoais da União Europeia (RGPD) e a Lei Brasileira de Proteção de Dados Pessoais (LGPD).*

Franco, P. A.(2020). *Lei Geral de Proteção de Dados Comentada*. Imperium Editora.

Kuner, C. (2019). The GDPR and International Organizations. *American Journal of International Law Unbound*, *114*, 15-19.

Moreira, S. (2023). O RGPD e a sua aplicação transfronteiriça: até onde estão protegidos os nossos dados pessoais? *Revista do CEJ*, (2), 357-380. https://doi.org/10.21814/uminho.ed.151.20

Osman, B. H. de & Soares, J. A. (2020). Transferência internacional de dados pessoais: a extraterritorialidade do RGPD europeu e seus impactos. Em M. Wachowicz (Eds.), *Proteção de dados pessoais em perspectiva: Lgpd e Rgpd na* ótica *do Direito comparado*. (pp. 563-593). Gedai.

Parlamento Europeu e Conselho da União Europeia. (1995). Diretiva 95/46/CE do Parlamento Europeu e do Conselho, de 24 de outubro de 1995, relativa à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados. *Jornal Oficial das Comunidades Europeias*, L 281, 31-50. https://eur-lex.europa.eu/legal-content/PT/Txt/Pdf/?uri=OJ:L:1995:281:Full

Parlamento Europeu e Conselho da União Europeia. (2016). Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados e que revoga a Diretiva 95/46/CE (Regulamento Geral sobre a Proteção de Dados). *Jornal Oficial da União Europeia*, L 119, 1-88. https://eur-lex.europa.eu/legal-content/PT/Txt/?uri=celex%3A32016R0679

Ramalho, P. J. N. (2020). *A adequação do sistema brasileiro de proteção de Dados pessoais ao regime das transferências europeu*. (Dissertação de mestrado). Universidade Nova de Lisboa. https://run.unl.pt/bitstream/10362/132234/1/Ramalho_2020.pdf

Rodrigues, E. H. K. (2016). O direito antitruste na economia digital: implicações concorrenciais do acesso a dados. (Dissertação de mestrado). Universidade de Brasília. https://repositorio.unb.br/bitstream/10482/20530/1/2016_EduardoHenriqueKruelRodrig ues.pdf

Schweighofer, E. (2017). Principles for US–EU Data Flow Arrangements. Em D. J. B. Svantesson & D. Kloza (Eds.), *Trans-Atlantic data privacy relations as a challenge for democracy* (pp. 27-47). Intersentia.

Tribunal de Justiça da União Europeia (2020, 16 de julho). *Data Protection Commissioner v. Facebook Ireland Ltd e Maximillian Schrems*, Processo C-311/18, Ecli:EU:C:2020:559. https://curia.europa.eu/juris/document/document.jsf?text=&docid=228677&pageIndex=0&doclang=PT

Tribunal de Justiça da União Europeia. (2015, 6 de outubro). *Maximillian Schrems v. Data Protection Commissioner*, Processo C-362/14, Ecli:EU:C:2015:650 https://eur-lex.europa.eu/legal-content/PT/Txt/Pdf/?uri=Celex:62014CJ0362

União Europeia. (2016). Carta dos Direitos Fundamentais da União Europeia. *Jornal Oficial da União Europeia*, C 202, 391-407. https://eur-lex.europa.eu/legal-content/PT/Txt/Pdf/?uri=Celex:12016P/Txt

Viola, M. (2019). *Transferência de dados entre Europa e Brasil: Análise da Adequação da Legislação Brasileira*. Instituto de Tecnologia & Sociedade do Rio. https://itsrio.org/wp-content/uploads/2019/12/Relatorio_UK_Azul_INTERACTIVE_Justificado.pdf

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 121-137

FUNDAMENTAÇÃO JURÍDICA E SISTEMAS DE IA: DA VALORAÇÃO À EXPLICABILIDADE CONTÍNUA

LEGAL FOUNDATIONS AND AI SYSTEMS: FROM VALUATION TO CONTINUOUS EXPLAINABILITY

Anna Carolina Borges Magalhães Bergamini, Gilberto Fernandes Taboada e Tomás Loureiro Soares

Universidade do Minho

Resumo

A crescente adoção de sistemas de Inteligência Artificial (IA) em contexto jurídico levanta questões pertinentes sobre a racionalidade das decisões e sobre a necessidade de assegurar a sua explicabilidade. A automatização dos processos decisórios desafia os princípios da fundamentação jurídica, que exigem transparência, coerência e possibilidade de recorrer das decisões. Este relatório examina a compatibilidade entre os ideais de fundamentação racional das decisões judiciais e a utilização de sistemas baseados em aprendizagem automática (*machine learning*), com especial atenção para a explicabilidade ao longo do ciclo de vida dos mesmos. Argumenta-se que a legitimidade de tais sistemas dependerá sempre da sua capacidade de manter uma relação coerente entre dados, modelos e justificações, permitindo a harmonia entre as decisões tomadas e aqueles que são os princípios nucleares de qualquer (ou quase todo o) ordenamento jurídico.

Palavras-chave

Fundamentação judicial, explicabilidad, inteligência artificial

Abstract

The growing adoption of Artificial Intelligence (AI) systems in the legal context raises pertinent questions about the rationality of decisions and the need to ensure their explainability. The automation of decision-making processes challenges the principles of legal reasoning, which demand transparency, coherence, and the possibility of appealing decisions. This report examines the compatibility between the ideals of rational reasoning in judicial decisions and the use of systems based on machine learning, with special attention to explainability throughout their lifecycle. It is argued that the legitimacy of such systems will always depend on their ability to maintain a coherent relationship between data, models, and justifications, allowing for harmony between the decisions made and the core principles of any (or almost every) legal system.

Keywords

Judicial reasoning, explainability, artificial intelligence

I. Introdução

A racionalidade é um elemento fundamental da decisão judicial moderna. O dever de fundamentação das decisões, consagrado em praticamente todos os ordenamentos jurídicos democráticos, traduz a exigência de que o juiz apresente razões compreensíveis, verificáveis e justificáveis para a sua tomada de posição — essa exigência cumpre uma dupla função: garantir a legitimidade democrática do poder jurisdicional e possibilitar o controlo público das decisões.

Nos últimos anos, a crescente utilização de ferramentas de Inteligência Artificial no mundo jurídico trouxe novas oportunidades e novos riscos para o

Direito — a promessa de eficiência, coerência e previsibilidade convive com o temor de perda de transparência, enviesamento algorítmico e perda da componente humana no processo decisório. Daqui, emergem duas dimensões críticas: a fundamentação racional das decisões e a explicabilidade dos sistemas de IA que participam, direta ou indiretamente, nesse processo. A dinâmica entre essas duas dimensões é o foco central deste relatório.

O objetivo é analisar de que modo a automatização das decisões judiciais pode (ou não) ser compatível com o ideal de racionalidade e fundamentação que estrutura o Direito. Sustenta-se que tal compatibilidade dependerá da explicabilidade contínua dos sistemas de IA ao longo do seu ciclo de uso — desde a conceção e treino até à implementação e monitorização — de modo a permitir a reconstrução da ratio subjacente a cada decisão.

A investigação desenvolve-se em cinco eixos principais: 1) caracterização da racionalidade jurídica e dos seus requisitos de fundamentação; 2) análise dos modelos de decisão automatizada e da sua lógica interna; 3) exame das abordagens de explicabilidade e dos desafios de aprendizagem contínua; 4) discussão crítica sobre os limites e condições de legitimidade da automatização no contexto jurídico; 5) casos e experiências concretas.

II. Fundamentação racional das decisões judiciais e valoração

A fundamentação racional das decisões judiciais é, historicamente, um dos pilares da racionalidade prática do Direito. Vários autores, entre os quais Robert Alexy (1989), destacam que o discurso jurídico é um caso especial do discurso prático racional, no qual a legitimidade das decisões depende da qualidade das razões apresentadas.

Na tradição do positivismo jurídico, a racionalidade era frequentemente associada à subsunção lógica: aplicar uma norma geral (premissa maior) a um caso concreto (premissa menor) para extrair uma conclusão. Contudo, a teoria contemporânea da argumentação jurídica reconhece que o raciocínio jurídico não é puramente dedutivo —envolve juízos de valor, ponderações casuísticas e escolhas interpretativas— elementos caraterísticos da natureza humana, que resistem à formalização total.

A teoria do discurso prático racional proposto por Alexy revela a racionalidade prática não como resultado em si, mas como processo de fundamentação que o atinge.

Considerando a existência de uma infinita quantidade de circunstâncias sociais que compõem a realidade humana, e que tais circunstâncias são fatos jurídicos em potência, ou seja, situações que podem ou não vir a ter relevância normativa para o mundo jurídico, percebe-se a impossibilidade de limitar a fundamentação racional das decisões judiciais unicamente através de precedentes lógicos.

Para Alexy, a decisão judicial não pode tornar-se um processo automático que sucede o resultado de premissas lógicas[1]. Em que pese o autor considerar a importância de um desfecho lógico a partir da premissa maior (norma jurídica válida) e uma premissa menor (fato comprovado), chamada por ele de justificação interna, tal situação não é suficiente. Isso se deve a algumas razões citadas pelo autor como a linguagem legal ser vaga, os possíveis conflitos entre normas, a hipótese de inexistência da norma jurídica para determinado fato e ainda as circunstâncias em que o juiz é obrigado a afastar a lei por ser incabível sua aplicação em determinado caso concreto. A justificação interna refere-se à lógica do silogismo, enquanto a justificação externa, também defendida pelo autor, valida as premissas adotadas na justificação interna através de fundamentação das próprias premissas.

Neste contexto, pensando nas premissas maior e menor percebe-se que ambas estão sujeitas a margem de interpretação. A adequação da norma, por vezes, não é unívoca e os fatos geralmente não atingem uma certeza absoluta, sendo constituídos a partir da produção de provas. Para que um raciocínio jurídico seja considerado sólido não é necessário a lógica em sentido próprio, mas a observação de casos reais, em um mundo em constante movimento, onde se infere a relevância da justificação externa defendida por Alexy, já que intenciona validar as premissas utilizadas na decisão.

Para operacionalizar a justificação externa cabe citar Atienza, que aborda que a construção da decisão judicial decorre de diversos tipos de argumentos, que formam um complexo para validar a razão[2]. O autor salienta a importância de ponderação de normas e princípios jurídicos no caso de conflitos entre eles, garantindo uma solução mais proporcional na aplicação no caso concreto. Ademais, ao final da obra o autor sugere três funções que a teoria da argumentação jurídica deveria cumprir, quais sejam, função teórica ou cognoscitiva (defende uma articulação do Direito com outros campos de conhecimento), função prática ou técnica (criação de sistemas jurídicos hábeis para inferência jurídica e auxílio do pensar jurídico) e função política ou moral (sugere uma perspetiva mais crítica e realista, que reconheça que nem sempre é possível alcançar plenamente a justiça material).

Segundo Taruffo, na maioria das estruturas processuais contemporâneas, a responsabilidade de justificar suas decisões é atribuída a todos os órgãos judiciais, e é frequente que esse requisito esteja consagrado em nível constitucional[3], conforme exemplificado pelo artigo 111.º, inciso 6, da Constituição italiana.

Da mesma forma, em Portugal o artigo 205.º, nº 1 da CRP prevê a necessidade de fundamentação das decisões dos tribunais, além de presentes os princípios da livre apreciação da prova, do contraditório e da fundamentação, consagrados no

1 Alexy, R. *A Theory of Legal Argumentation* (Oxford: Clarendon Press, 1989), p. 221.

2 Atienza, M. *As Razões do Direito. Teorias da Argumentação Jurídica*, trad. M. C. G. Cupertino (São Paulo: Landy Editora, 2003), pp. 224-225.

3 Taruffo, M. *Simplemente la verdad. El juez y la construcción de los hechos* (Madrid: Marcial Pons, 2010), pp. 134-135.

Código de Processo Penal (arts. 127.º, 327.º e 374.º, nº 2) e no Código de Processo Civil (art. 607.º).

Por sua vez, como bem aponta Taruffo, em sistemas jurídicos alternativos, como o da Inglaterra, não existe obrigação explícita para fornecer justificativa, no entanto, existe uma prática judicial bem estabelecida nesse contexto. Além disso, em todos esses casos, o juiz deve articular a lógica por trás de sua decisão, delineando os fundamentos sobre o qual deve ser considerado válido e racionalmente justificado[4].

Os fundamentos de uma decisão não devem conter mera derivação lógica de suposições ou teoremas probabilísticos, devendo o juiz, em seu grau de discricionariedade, realizar a apreciação das provas através da racionalidade de valoração, evitando que se torne um ato de arbitrariedade.

Neste sentido, Taruffo destaca que elucidar os fatos implica em explicar, na forma de um raciocínio justificativo, o raciocínio lógico que elege uma eficácia particular a cada meio de prova e que, com base nisso, fundamenta a preferência pela escolha da hipótese com base na premissa de que, dada as opções disponibilizadas em questão, essa possui um maior grau de corroboração lógica.

Ainda segundo o autor a noção de motivação como fundamentação racional do julgamento, encontra um pilar no quesito de controle derivado da discricionariedade do juiz na avaliação da prova, cumprindo o papel de regular essa discricionariedade, obrigando o juiz a racionalizar suas próprias decisões e permitindo uma avaliação subsequente dessas decisões, durante e após o processo judicial[5].

A fundamentação racional, portanto, exige coerência interna (ausência de contradições), correspondência entre normas e factos (adequação empírica) e justificação discursiva (possibilidade de defesa pública das razões adotadas). Trata-se de uma racionalidade comunicativa e justificativa, mais do que puramente instrumental.

No plano normativo, o dever de fundamentar as decisões traduz-se em obrigações de clareza, completude e motivação suficiente. Isso significa que o julgador deve indicar as normas aplicáveis, os factos relevantes e o caminho lógico que conduz da análise desses elementos à conclusão final.

Com a introdução da automatização, esta estrutura enfrenta um desafio: quem ou o que fundamenta a decisão quando parte do processo é mediado por algoritmos? Se a racionalidade jurídica se apoia na justificação discursiva, como pode ser preservada quando a decisão resulta de operações de cálculo ou de modelos de aprendizagem estatística?

A atuação do juiz a partir da aplicação do Direito pode ser influenciada por fragilidades inerentes ao ser humano; por outro lado, é justamente a presença humana que não se limita ao uso de regras para resolução do litígio, mas também sopesa caminhos para solução, podendo recorrer a outras fontes do Direito, tais como princípios, jurisprudência, analogias e costumes.

4 Taruffo, *Simplemente la verdad*…, pp. 138-139.

5 Taruffo, M. *La prueba de los hechos* (Madrid: Trotta, 2005), p. 435.

A fundamentação de uma decisão judicial elaborada através da IA atualmente não é capaz de tangenciar uma decisão fundamentada racionalmente por um juiz, ou seja, onde exista a expressão de uma consciência humana que garanta a compreensão do contexto, aplicação ético-normativa e ponderação de valores.

III. Racionalidade algorítmica e automatização

A difusão de sistemas de LLM (*Large Language Model*, Modelo de Linguagem Grande) para acesso geral ao público através da Rede Mundial de Computadores-Internet, tais como Chat GPT4 (lançado em novembro de 2022 pela Open AI), Grok 4 (lançado em julho de 2025 pela xAI) e Gemini (lançado em março de 2023 pela Google) tem feito surgir discussões sobre inteligência artificial e possibilidades de uso dessa tecnologia em áreas onde até pouco tempo não se imaginava haver a possibilidade de substituição do trabalho humano.

1. Inteligência artificial

A expressão «inteligência artificial» pode assumir diversos significados, todos normalmente relacionados com a capacidade de uma máquina apresentar respostas a problemas de forma semelhante a um ser humano.

Diversos autores já tentaram definir a expressão, com maior ou menor grau de sucesso. Ertel[6] cita duas definições interessantes: John McCarthy, pioneiro das pesquisas em IA, definiu o objetivo da IA da seguinte forma em 1955: «O objetivo da IA é desenvolver máquinas que se comportem como se fossem inteligentes». A definição é vaga o suficiente para admitir como IA qualquer máquina que «pareça» se comportar como um humano, sem estabelecer a necessidade de suporte de um sistema de computação.

Elaine Rich em 1983 definiu da seguinte forma: «Inteligência Artificial é o estudo de como fazer computadores executar tarefas que, no presente momento, pessoas fazem melhor». Essa é considerada elegante por Ertel, por definir de forma sucinta o que os cientistas tentam fazer desde o início das pesquisas nesse campo do conhecimento.

O tema de inteligência artificial é tão importante que o Congresso dos Estados Unidos da América publicou o *National Defense Authorization Act of 2019*, que conta com uma série de regras e metas a serem cumpridas no desenvolvimento de sistemas dessa natureza. Esse ato do Congresso conta com a seguinte definição de IA[7]:

1. Qualquer sistema artificial que execute tarefas sob circunstâncias diversas e imprevisíveis sem vigilância humana relevante, ou que pode aprender a partir da experiência e melhorar seu desempenho quando apresentado a um conjunto de dados;

6 Ertel, W. *Introduction to Artificial Intelligence*, 2.ª ed. (Cham: Springer, 2017), pp. 1-3.

7 Congress of the United States. *H.R.5515 – John S. McCain National Defense Authorization Act for Fiscal Year 2019*, 2018, pp. 62-63. https://www.congress.gov/bill/115th-congress/house-bill/5515/text

2. Um sistema artificial desenvolvido em software de computador, *hardware* físico ou outro contexto que execute tarefas que demandem percepção, cognição, planejamento, aprendizado, comunicação ou ação física semelhantes à de um ser humano;

3. Um sistema artificial projetado para pensar ou agir como um humano, inclusive arquiteturas cognitivas e redes neurais;

4. Um conjunto de técnicas, inclusive aprendizado de máquina (*machine learning*) que é projetado para se assemelhar a uma tarefa cognitiva;

5. Um sistema artificial projetado para agir racionalmente, inclusive um agente inteligente de software ou robô que atinja objetivos usando percepção, planejamento, raciocínio, aprendizado, comunicação, tomada de decisões e ação.

2. Ciclo de vida de desenvolvimento de sistemas de IA

Tradicionalmente, o desenvolvimento de um sistema de informação envolve as seguintes etapas:

- Análise de requisitos: nesta fase os desenvolvedores tratam de entender o problema que se quer resolver. Normalmente isso é feito através de estudos do modelo de negócio que se quer tratar e entrevistas com os chamados *stakeholders* — pessoas que conhecem o modelo de negócio e eventualmente se tornarão usuários do sistema. Nesta fase se define o que o sistema fará.
- Projeto: nesta fase se planeja como serão implementadas as funcionalidades definidas no passo de análise de requisitos. Nesta fase então se definem *frameworks* a serem utilizados, linguagens de programação, sistemas de gerenciamento de bases de dados, tecnologias de comunicação, entre outros.
- Construção: é a fase de codificação do sistema, ou seja, de programação propriamente dita, com eventual integração entre diferentes tecnologias que foram definidas na fase de projeto.
- Teste: nesta fase o sistema pode estar parcialmente ou totalmente construído. Elabora-se um plano de teste que contem dados fictícios, funções a serem testadas e os resultados esperados para cada teste. As funcionalidades testadas são aquelas definidas na análise de requisitos.
- Produção: uma vez que o sistema esteja totalmente construído e os testes tenham dado resultados satisfatórios o sistema entra em fase de produção, i.e., entra em funcionamento efetivo em proveito dos *stakeholders*.

Não há consenso entre desenvolvedores de sistemas de informação em como implementar essas etapas. Há metodologias que usam o processo de «cascata» (*waterfall*), em que somente se passa a próxima etapa quando a etapa anterior está totalmente esgotada[8]. Outros especialistas preferem modelos ditos «ágeis» (*agile*), em que as funcionalidades do sistema são implementadas aos poucos, e as etapas de análise de requisitos, projeto, construção, teste e produção se

8 Royce, W. «Managing the Development of Large Software Systems». *Proceedings of IEEE WESCON*, 1970.

sucedem rapidamente num modelo conhecido como «espiral», onde a cada ciclo são adicionadas funcionalidades ao sistema.

O ciclo de vida de um sistema de IA difere do modelo clássico de desenvolvimento de sistemas por causa da importância da fase de aprendizado do sistema. Há a necessidade de coleta de dados, análise crítica e aprendizado por parte do sistema. Uma vez que o sistema tenha «aprendido» o que é necessário para seu funcionamento ele entra em produção, mas durante essa fase naturalmente o sistema continua aprendendo com novos dados que são inseridos e demandas dos usuários.

O Ciclo de vida de um sistema de IA tipicamente envolve as seguintes fases:

- Projeto: envolve definição do problema a ser resolvido, coleta de dados de treinamento, preparação dos dados e seleção do modelo de IA e projeto de arquitetura de sistema.
- Desenvolvimento: nesta etapa o sistema é alimentado com os dados definidos na etapa anterior — é o chamado «treinamento» do modelo. Passa-se a seguir à avaliação e teste do modelo, e finalmente à etapa de refinamento iterativo.
- Início de produção e manutenção: nesta etapa o sistema entra em funcionamento para os usuários. Deve-se fazer a partir daí um monitoramento constante do comportamento do sistema, e eventualmente promover manutenção no modelo e *feedback* (relatórios de atividades) aos *stakeholders*.

3. Automatização das decisões judiciais

A tentativa de automatizar decisões judiciais não surge com a ascensão dos LLM. Desde a década de 1980, sistemas *rule-based* procuravam modelar o raciocínio jurídico através de regras lógicas formalizadas («se-então»). Esses sistemas apresentavam elevada explicabilidade, pois cada inferência era rastreável. Contudo, a sua rigidez e incapacidade de lidar com ambiguidades linguísticas e contextuais, naturalmente, limitaram a sua aplicação. No domínio da IA, o machine learning constitui o seu núcleo técnico fundamental. Nos modelos supervisionados, o sistema aprende com exemplos rotulados, aproximando categorias ou valores fornecidos previamente por humanos. Já nos modelos não supervisionados, a aprendizagem decorre de forma autónoma, mediante a identificação de estruturas e relações latentes nos dados, sem rótulos definidos à partida. Esta distinção é essencial para compreender tanto o potencial como as limitações dos sistemas aplicados ao Direito.

Com o avanço do machine learning, a ênfase deslocou-se da programação explícita de regras para a aprendizagem a partir de dados. Modelos conexionistas, como as redes neuronais, podem identificar padrões complexos em grandes volumes de decisões judiciais, prevendo resultados com alto grau de acerto. Todavia, esses modelos introduzem uma nova forma de racionalidade: a racionalidade algorítmica, baseada em correlações empíricas, não em justificações normativas.

A racionalidade algorítmica tende a privilegiar eficiência preditiva em detrimento da justificação interpretativa. Em contextos jurídicos, essa diferença é

crucial: uma decisão que é estatisticamente correta pode, ainda assim, ser juridicamente injustificável se não puder ser explicada nos termos do ordenamento jurídico e dos valores constitucionais, nomeadamente no que diz respeito a ponderações sobre o caso concreto.

Para além disso, a automatização introduz problemas de valoração: critérios aprendidos a partir de dados históricos podem reproduzir vieses sociais, de género ou raciais, transferindo injustiças do passado para o futuro com aparência de objetividade.

Assim, a automatização das decisões judiciais coloca em confronto dois ideais distintos: o da racionalidade formal-computacional e o da racionalidade prática-discursiva do Direito.

IV. Explicabilidade ao longo do ciclo de uso

O conceito de explicabilidade (ou explainability) surge precisamente como resposta a essa tensão. Enquanto a fundamentação racional visa a justificação perante a comunidade jurídica e social, a explicabilidade procura tornar compreensível o funcionamento interno de modelos algorítmicos para os seus utilizadores humanos.

O *Guia para a Inteligência Artificial* da Agência para a Modernização Administrativa define o conceito de explicabilidade da seguinte forma:

> A Explicabilidade é o primeiro princípio no pilar da Transparência. Explicabilidade é a capacidade de um sistema de IA articular a lógica por detrás das suas decisões e resultados de forma compreensível para os utilizadores e partes interessadas. Não basta que um sistema seja transparente na sua conceção e documentação, precisa também de ser inteligível na sua operação diária para construir confiança e permitir uma supervisão informada. Esta capacidade de fornecer um caminho de decisão rastreável é essencial, pois não só desmistifica o comportamento da IA, como também se torna uma ferramenta de diagnóstico fundamental para identificar e corrigir erros[9].

A literatura técnica distingue entre explicabilidade global (compreensão geral da estrutura do modelo) e explicabilidade local (justificação de decisões específicas). Métodos como LIME, SHAP e counterfactual explanations permitem, por exemplo, identificar quais variáveis mais influenciaram uma decisão concreta — contudo, tais explicações nem sempre correspondem a uma fundamentação jurídica no sentido estrito, pois descrevem relações causais ou estatísticas, não razões normativas.

Para o Direito, a explicabilidade assume um duplo significado: epistemológico, porque assegura a possibilidade de compreender e auditar o raciocínio do sistema; e normativo, porque garante o cumprimento do dever de fundamentação e o direito das partes de conhecer as razões da decisão. O artigo 14.º do *AI*

9 Mosaico, *Guia Prático para a Inteligência Artificial Ética,* 2025, pp. 38-39. https://bo.mosaico.gov.pt/api/assets/mosaico-site/ff6944f0-8046-46ef-8ae0-36126b5f3b1f

Act introduz o requisito de human oversight, que implica a existência de mecanismos de explicabilidade proporcionais ao contexto de utilização. No domínio da justiça, isto significa que a explicação deve ser funcionalmente adequada às exigências da fundamentação jurídica: o utilizador humano deve compreender as razões que conduziram a um *output* do sistema e poder contrariá-lo em função das normas e dos princípios aplicáveis ao caso.

A exigência de explicabilidade, porém, não se esgota na interface final do sistema - deve acompanhar todo o ciclo de vida da IA desde a conceção, treino e validação até à implementação e monitorização contínua. A cada etapa, surgem riscos específicos:

- Na fase de conceção, escolhas de modelo e variáveis determinam as dimensões da realidade jurídica que serão representadas.
- Durante o treino, a seleção e rotulagem de dados influenciam a aprendizagem e podem introduzir viés.
- Na implementação, o contexto de uso define a interpretação dos resultados (decisão assistida ou substitutiva).
- Na monitorização, o modelo deve ser auditado, corrigido e eventualmente treinado de forma contínua, para evitar deriva de desempenho ou desvio normativo.

A manutenção da explicabilidade ao longo deste ciclo constitui condição essencial para a legitimidade e responsabilidade jurídica das decisões automatizadas.

Gunning e Aha apresentam o programa de desenvolvimento de Inteligência Artificial Explicável (XAI, na sigla em inglês) da DARPA (Defense Advanced Research Projects Agency, USA)[10]. Segundo os autores, há modelos de IA que obtiveram muito sucesso em áreas como transporte, segurança, medicina, finanças e defesa, mas que não são capazes de explicar aos usuários suas decisões.

Em oposição aos sistemas de IA ditos «opacos» (*blackbox*), que não oferecem detalhes aos usuários sobre o processo de «raciocínio» do modelo, há uma demanda por sistemas de IA «transparentes», que explicam em linguagem compreensível por seres humanos os passos que levaram o modelo à resposta a um problema.

Tal funcionalidade é essencial para diversos tipos de aplicação pelas mais diversas razões. Tipicamente muitos usuários ou mesmo pessoas que indiretamente podem beneficiar ou sofrer as consequências de uma decisão de IA (por exemplo, de uma sentença judicial) vão querer saber o porquê daquela decisão para poderem confrontá-la em outro ambiente.

V. Casos e experiências concretas

A análise de casos concretos permite observar como os ideais de fundamentação racional e explicabilidade se manifestam (ou não) quando são aplicados sistemas reais de inteligência artificial no domínio jurídico — apresentamos, de

10 Gunning, D., Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, vol. 40, n.º 2 (2019), p. 44.

seguida, dois casos paradigmáticos, um nos EUA e um na Europa, que ilustram os benefícios e os riscos da automatização de decisões judiciais.

1. O sistema COMPAS

O Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) é um sistema algorítmico de apoio à decisão judicial desenvolvido pela empresa norte-americana Northpointe Inc., cujo objetivo é avaliar o risco de reincidência criminal com base em dados recolhidos sobre o histórico dos arguidos — como idade, antecedentes, emprego, contexto social e respostas a determinados questionários.

Em 2016, uma investigação do ProPublica revelou que o COMPAS apresentava um viés racial sistemático: pessoas afrodescendentes eram classificadas com risco elevado de reincidência em proporção significativamente maior relativamente a pessoas caucasianas com registos criminais idênticos[11]. Além disso, os critérios de ponderação do modelo eram privados e inacessíveis, impossibilitando o escrutínio público — o dever de fundamentação explica-se pela necessidade de justificação do exercício do poder estadual, da rejeição do segredo nos atos do Estado, da necessidade de avaliação dos atos estaduais, aqui se incluindo a controlabilidade, previsibilidade, fiabilidade e a confiança nos atos do Estado[12].

Segundo Hildebrandt, este tipo de opacidade técnica constitui uma forma de «governança algorítmica» que ameaça princípios fundamentais do Estado de Direito, ao deslocar o locus da justificação racional para uma esfera meramente tecnocientífica, fora do debate jurídico[13]. A ausência de transparência rompe com o ideal de racionalidade comunicativa formulado por Habermas [14], no qual a legitimidade depende da possibilidade de justificar racionalmente as decisões perante os cidadãos.

Entre eficiência e racionalidade jurídica

A defesa do COMPAS baseou-se na alegada eficiência preditiva do modelo — ou seja, na sua capacidade de correlacionar as variáveis recolhidas com a reincidência futura dos arguidos. No entanto, a legitimidade da decisão judicial não reside apenas na correção empírica, mas na fundamentação discursiva das razões normativas que sustentam a decisão[15]. O sistema, por mais preciso que pudesse ser, não oferecia razões compreensíveis no sentido jurídico, mas apenas outputs numéricos resultantes de correlações estatísticas.

11 Angwin, J., Larson, J., Mattu, S., Kirchner, L. «Machine Bias: There's Software Used Across the Country to Predict Future Criminals». *ProPublica*. 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

12 Canotilho, J. J. G., Moreira, V. *Constituição da República Portuguesa Anotada*, vol. II, 4.ª ed. (Coimbra: Coimbra Editora, 2010), p. 527.

13 Hildebrandt, M. *Smart Technologies and the End(s) of Law* (Cheltenham: Edward Elgar, 2015).

14 Habermas, J. *Teoria do Agir Comunicativo*, vol. I. (Lisboa: Edições 70, 1987).

15 Alexy, *A Theory of Legal Argumentation*, pp. 230-231.

Doshi-Velez e Kim denominam este tipo de desempenho «eficácia sem interpretabilidade»: o modelo pode produzir previsões úteis, mas a falta de interpretabilidade torna-as epistemicamente opacas[16]. Em termos jurídicos, essa opacidade converte-se num défice de fundamentação, contrariando o dever de motivação das decisões judiciais previsto em praticamente todos os ordenamentos constitucionais democráticos. No contexto norte-americano, o caso *State v. Loomis* (2016) levou o problema aos tribunais. O Supreme Court of Wisconsin reconheceu o risco de violação do *due process of law* (devido processo legal), mas manteve a admissibilidade do COMPAS, sob a condição de que fosse utilizado apenas como instrumento auxiliar e não como o único determinante da pena.

Numa perspetiva europeia, um sistema análogo enfrentaria barreiras jurídicas mais robustas: o artigo 47.º da Carta dos Direitos Fundamentais da União Europeia (CDFUE) e o artigo 205.º da Constituição da República Portuguesa (CRP) consagram o direito à fundamentação das decisões judiciais. Além disso, o mais recente Regulamento Europeu de Inteligência Artificial (*AI Act*, Regulamento UE 2024/1689) classifica as aplicações de IA no domínio judicial como «sistemas de alto risco», exigindo níveis elevados de transparência, documentação e explicabilidade. Como sublinha Miller, a explicabilidade não é apenas uma característica técnica, mas uma condição relacional de confiança na máquina: o output só é legítimo se os destinatários puderem compreender e avaliar as razões que o sustentam[17]. Assim, no plano normativo, a transparência algorítmica deve ser vista como extensão contemporânea do princípio jurídico da motivação — não apenas para cumprir formalidades, mas para assegurar controlo público e *accountability*.

O caso COMPAS revela o conflito estrutural entre a racionalidade algorítmica e a racionalidade jurídica: a primeira baseia-se em correlações e probabilidades, enquanto a segunda exige justificação discursiva, motivação e possibilidade de contraditório. Os sistemas algorítmicos operam num plano de informação sintática, enquanto o Direito opera num plano de significação semântica e valorativa — o problema do COMPAS, portanto, não é apenas técnico, mas também ontológico — o que está em causa é a tradução de razões jurídicas em razões computacionais, e vice-versa[18].

2. Projeto AI4Justice

O projeto AI4Justice, desenvolvido no contexto europeu, visa explorar o potencial da IA para analisar grandes volumes de jurisprudência, organizar processos e apoiar decisões judiciais, mantendo sempre o juiz como decisor final - um exemplo clássico de abordagem human-in-the-loop: ao contrário de sistemas totalmente automatizados como o COMPAS, o AI4Justice não substitui a decisão humana, mas fornece informações estruturadas, resumos de precedentes e

16 Doshi-Velez, F., Kim, B. *Towards A Rigorous Science of Interpretable Machine Learning*, 2017.

17 Miller, T. «Explanation in artificial intelligence: Insights from the social sciences». *Artificial Intelligence*, 267 (2019).

18 Floridi, L. *The Ethics of Information* (Oxford: Oxford University Press, 2013).

sugestões baseadas em padrões estatísticos, permitindo aos magistrados tomar decisões mais informadas e consistentes.

Do ponto de vista da racionalidade jurídica, o AI4Justice apresenta vantagens claras: aumenta a eficiência, reduz a carga de trabalho e contribui para a uniformidade na aplicação do direito. No entanto, a utilização do sistema evidencia desafios relacionados com a fundamentação racional das decisões. Embora o algoritmo possa indicar quais precedentes ou normas são relevantes para um caso, não gera uma justificação jurídica completa, sendo sempre necessária a interpretação humana para transformar os dados em razões válidas e compreensíveis juridicamente.

Neste sentido, a integração de mecanismos de explicabilidade (XAI) ao longo do ciclo de vida do sistema mostra-se essencial. Cada sugestão produzida pelo AI4Justice deve poder ser compreendida, auditada e contestada, garantindo que o processo de decisão permanece transparente e legítimo. Esta abordagem reflete diretamente os princípios do *AI Act* da União Europeia[19], que exige supervisão humana e documentação detalhada para sistemas de IA de alto risco, como os utilizados em contexto judicial. Mesmo em sistemas supervisionados como o AI4Justice, subsiste o risco de ancoragem cognitiva - a recomendação algorítmica tende a funcionar como referência inicial do raciocínio, influenciando o juiz de forma subtil e diminuindo a independência crítica da análise. Quando a saída do modelo é apresentada com aparente fiabilidade estatística, o utilizador humano pode ser levado a confiar excessivamente nela, ajustando o raciocínio jurídico ao «ponto de partida» sugerido pelo sistema.

Em termos críticos, o AI4Justice exemplifica a tensão entre eficiência algorítmica e fundamentação racional: o sistema melhora a rapidez e a consistência das decisões, mas não substitui a necessidade de justificação discursiva e responsabilidade humana. A sua implementação reforça a importância de políticas institucionais que assegurem auditoria contínua, formação dos utilizadores e monitorização do desempenho, evitando que a automatização leve a decisões opacas ou não justificáveis.

Em suma, o AI4Justice representa um caso paradigmático de como a IA pode apoiar o processo decisório judicial, respeitando os limites impostos pela racionalidade jurídica e pela necessidade de explicabilidade — constitui um exemplo concreto de um benefício equilibrado com precauções éticas e jurídicas, ao não depositar todo o peso da decisão num conjunto de operações matemáticas e variáveis estatísticas.

VI. Discussão crítica

A integração da inteligência artificial no processo judicial suscita questões centrais sobre racionalidade, legitimidade e explicabilidade. A automatização não se reduz a uma questão técnica: envolve um repensar da própria epistemologia jurídica, da normatividade do discurso judicial e das estruturas institucionais que sustentam o exercício do Direito.

19 European Commission. *Regulation (EU) 2024/1689 – Artificial Intelligence Act*, art. 14º.

1. Dimensão epistemológica: racionalidade algorítmica e racionalidade jurídica

A IA opera segundo uma racionalidade algorítmica, baseada em correlação estatística e inferência probabilística, distinta da racionalidade argumentativa que fundamenta o Direito. Como defende Hildebrandt, esta transição representa uma mudança de paradigma: da hermenêutica da norma para a previsão do comportamento[20]. O risco está em substituir a justificação discursiva pela eficácia preditiva.

O caso COMPAS ilustra esse desvio: decisões judiciais condicionadas por um modelo opaco, sem explicabilidade adequada e sem possibilidade de reconstruir o raciocínio subjacente[21]. Já o AI4Justice adota uma racionalidade híbrida, conjugando processamento algorítmico com supervisão humana, procurando manter a coerência interpretativa e a rastreabilidade da decisão. Esta combinação mostra que a automatização pode coexistir com a racionalidade jurídica, desde que o sistema seja concebido como instrumento de apoio e não como substituto da deliberação.

2. Dimensão normativa: legitimidade e justificação pública

A automatização das decisões judiciais exige um quadro normativo que assegure legitimidade discursiva e responsabilidade pública. Como sustenta Habermas, a validade de uma decisão depende da possibilidade de justificar racionalmente os seus fundamentos perante todos os afetados[22]. Um sistema opaco, mesmo que eficaz, viola este princípio. A integração da IA em tarefas judiciais é enquadrada pelo Regulamento Europeu da Inteligência Artificial (Regulamento UE 2024/1689), que classifica estes sistemas como de alto risco e exige documentação técnica, mecanismos de gestão de risco e supervisão humana robusta (arts. 6.º a 29.º). Complementarmente, o art. 22.º do RGPD limita decisões exclusivamente automatizadas, impondo salvaguardas adicionais quando estejam em causa direitos ou interesses relevantes. Em Portugal, aplicam-se ainda as garantias constitucionais do processo justo e equitativo, previstas nos arts. 20.º e 32.º da CRP. Acresce ainda a Carta Portuguesa de Direitos Humanos na Era Digital (Lei 27/2021), que no art. 9.º impõe deveres reforçados de transparência e auditabilidade sempre que a Administração Pública recorra a sistemas automatizados. A explicabilidade, neste contexto, não é um requisito técnico acessório: é uma garantia jurídica de legitimidade, que permite o escrutínio e a contestação. Assim, a fundamentação racional e a explicabilidade não são dimensões totalmente distintas, mas partes de um mesmo imperativo normativo: o de tornar o poder de decidir juridicamente justificável.

20 Hildebrandt, *Smart Technologies*…

21 Doshi-Velez e Kim, *Towards A Rigorous Science*…

22 Habermas, *Teoria do Agir Comunicativo*.

3. Dimensão institucional: governança e responsabilidade algorítmica

Do ponto de vista institucional, a adoção de IA no sistema judicial implica repensar mecanismos de governança algorítmica[23]. A delegação parcial de tarefas decisórias a sistemas automáticos requer novas formas de auditoria, formação interdisciplinar e mecanismos de correção contínua. A experiência do AI4Justice demonstra que a criação de equipas mistas — juízes, informáticos e juristas especializados em ética e regulação — é essencial para garantir que os algoritmos se mantêm alinhados com princípios de proporcionalidade e justiça.

A responsabilidade, por sua vez, deve permanecer claramente atribuída a agentes humanos — a grande questão será sobre quem recai a responsabilidade ao certo: quem programa ou quem aplica? Como sublinha Alexy, a decisão jurídica só é legítima enquanto puder ser inserida num discurso racional e imputável[24]. Mesmo num contexto automatizado, a autoridade de decidir deve manter-se ancorada na racionalidade e na ética do julgador, não na opacidade do código. Do ponto de vista legal, a questão da responsabilidade quando uma recomendação algorítmica contribui para uma decisão injusta permanece em aberto. O *AI Act* define requisitos de segurança e supervisão, mas não estabelece um regime substantivo de responsabilidade civil. Em Portugal, vigora o regime geral da responsabilidade do Estado e dos titulares de cargos públicos (Lei nº 67/2007), segundo o qual o juiz continua responsável pela decisão final — todavia, esta solução mostra-se insuficiente quando a influência algorítmica é estrutural ou difícil de detetar, criando zonas de responsabilidade difusa que exigem reflexão normativa aprofundada.

VII. Conclusão

A investigação desenvolvida permite afirmar que a automatização das decisões judiciais não constitui apenas uma evolução tecnológica, mas uma verdadeira reconfiguração epistemológica e institucional do Direito. Este fenómeno desafia a própria ideia de racionalidade jurídica, historicamente ancorada na argumentação, na interpretação e na fundamentação discursiva, ao introduzir formas de racionalidade algorítmica baseadas em modelos preditivos e inferências estatísticas.

A análise comparativa entre o COMPAS e o AI4Justice mostrou de modo exemplar como esta transição pode seguir trajetórias divergentes. O COMPAS, amplamente criticado pela sua opacidade e enviesamento, representa um paradigma de automatização sem explicabilidade suficiente, no qual a decisão é sustentada por inferências não acessíveis nem verificáveis pelos sujeitos afetados. Em contraste, o AI4Justice demonstra que é possível integrar a IA no processo judicial sem abdicar da racionalidade e da legitimidade discursiva, adotando

23 Floridi, L. *The Logic of Information: A Theory of Philosophy as Conceptual Design* (Oxford: Oxford University Press, 2020).

24 Alexy, *A Theory of Legal Argumentation*.

mecanismos de supervisão humana e rastreabilidade ao longo de todo o ciclo de uso do sistema.

Esta diferença traduz, em termos normativos, duas conceções distintas de justiça: uma instrumental, centrada na eficiência e previsibilidade, e outra deliberativa, orientada pela transparência, responsabilidade e justificação pública. No contexto europeu, é esta segunda conceção que se impõe, em consonância com os princípios do *AI Act*, que reconhecem a necessidade de mecanismos de explicabilidade técnica e jurídica para sistemas de alto risco, especialmente no domínio judicial.

Contudo, a mera introdução de requisitos legais e técnicos não é suficiente. A verdadeira legitimidade da automatização exige uma transformação cultural e institucional mais ampla: é necessário desenvolver competências interdisciplinares que permitam aos profissionais do Direito compreender, auditar e questionar os sistemas de IA que utilizam. Do mesmo modo, os engenheiros e cientistas de dados envolvidos na conceção destes sistemas, devem ver também maior investimento na sua formação em ética e raciocínio jurídico, de modo a incorporar valores como imparcialidade, proporcionalidade e contraditório nos próprios modelos de decisão[25].

Além disso, a introdução de IA nos tribunais requer mecanismos permanentes de auditoria algorítmica — estruturas que garantam a verificação periódica da equidade, da precisão e da consistência dos sistemas, bem como a possibilidade de correção de desvios. Estas auditorias devem ser transparentes e participadas, envolvendo não apenas técnicos, mas também profissionais do ramo jurídico e, idealmente, a sociedade civil. Tal abordagem está em linha com as propostas de Hildebrandt, que defende um modelo de «governança algorítmica reflexiva»[26], em que os sistemas tecnológicos permanecem sujeitos a um escrutínio jurídico e público constante.

Por conseguinte, a compatibilização entre IA e racionalidade jurídica depende de um equilíbrio delicado: utilizar a capacidade técnica das máquinas para ampliar a compreensão humana, sem nunca substituir o julgamento humano pela opacidade algorítmica. A IA deve ser concebida como instrumento de apoio ao raciocínio jurídico, não como substituto da autoridade judicial.

Em síntese, a legitimidade do futuro digital do Direito dependerá de três condições centrais:

- Supervisão e imputabilidade humana efetiva, assegurando que a responsabilidade última pelas decisões permanece dos magistrados.
- Explicabilidade técnica e discursiva, permitindo que o raciocínio algorítmico seja compreensível e juridicamente justificável.
- Governança institucional transparente, que promova auditoria, formação e revisão contínua dos sistemas.

Se estas condições forem respeitadas, a inteligência artificial poderá tornar-se um aliado da racionalidade jurídica, promovendo maior coerência, eficiência

25 Floridi, *The Logic of Information*…

26 Hildebrandt, *Smart Technologies*…

e equidade nas decisões, sem sacrificar os princípios fundamentais do Estado de Direito. Assim, mesmo num contexto de transformação tecnológica profunda, o Direito poderá continuar a cumprir a sua missão essencial: produzir decisões humanas, racionais e justificáveis, agora com o apoio de tecnologias capazes de ampliar, mas nunca substituir, a razão prática que o sustenta.

Referências bibliográficas

Alexy, R. *A Theory of Legal Argumentation*. Oxford: Clarendon Press, 1989.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. *ProPublica*, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Atienza, M. *As Razões do Direito – Teorias da Argumentação Jurídica*, trad. M. C. G. Cupertino. São Paulo: Landy Editora, 2003.

Canotilho, J. J. G., Moreira, V. *Constituição da República Portuguesa Anotada*, vol. II, 4.ª ed. Coimbra: Coimbra Editora, 2010.

Congress of the United States. *H.R.5515 – John S. McCain National Defense Authorization Act for Fiscal Year 2019*. https://www.congress.gov/bill/115th-congress/house-bill/5515/text

Doshi-Velez, F., Kim, B. «Towards a Rigorous Science of Interpretable Machine Learning». *ArXiv*, 2017. https://doi.org/10.48550/arXiv.1702.08608

Ertel, W. *Introduction to Artificial Intelligence*, 2.ª ed. Cham: Springer, 2017.

European Commission. *Regulation (EU) 2024/1689 – Artificial Intelligence Act*. 2014. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Floridi, L. *The Ethics of Information*. Oxford: Oxford University Press, 2013.

Floridi, L. *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford: Oxford University Press, 2020.

Gunning, D., Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, vol. 40, n.º 2 (2020): 1-58. https://doi.org/10.1609/aimag.v40i2.2850

Habermas, J. *Teoria do Agir Comunicativo*. Lisboa: Edições 70, 1987.

Hildebrandt, M. *Smart Technologies and the End(s) of Law*. Cheltenham: Edward Elgar, 2015.

Miller, T. «Explanation in Artificial Intelligence: Insights from the Social Sciences». *Artificial Intelligence*, 267 (2019): 1-38.

Mosaico. *Guia Prático para a Inteligência Artificial Ética*, 2025. https://bo.mosaico.gov.pt/api/assets/mosaico-site/ff6944f0-8046-46ef-8ae0-36126b5f3b1f

Royce, W. «Managing the Development of Large Software Systems». *Proceedings of IEEE WESCON*, 1970.

Taruffo, M. *La prueba de los hechos*. Madrid: Trotta, 2025.

Taruffo, M. *Simplesmente la verdad. El juez y la construcción de los hechos*. Madrid: Marcial Pons, 2010.

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 138-155

CUESTIONES EN TORNO A LA APTITUD DE LA INTELIGENCIA ARTIFICIAL PARA ACELERAR EL PROCESO PENAL EN SUPUESTOS DE MULTIRREINCIDENCIA DESDE LA FISCALÍA DE BARCELONA[1]

ISSUES SURROUNDING THE SUITABILITY OF ARTIFICIAL INTELLIGENCE TO ACCELERATE CRIMINAL PROCEEDINGS IN CASES OF PERSISTENT OFFENDING FROM THE PERSPECTIVE OF THE BARCELONA PUBLIC PROSECUTOR'S OFFICE

Laura Echarri Nicolás

Graduada en Derecho, máster en Derecho digital y de la inteligencia artificial, experta en inteligencia artificial y estudiante de Máster en Abogacía y Procura

Diego Fierro Rodríguez

Letrado de la Administración de Justicia en el Ministerio de Justicia (España)

1 Este trabajo se enmarca en el Proyecto I+D+i de generación de conocimiento y fortalecimiento científico y tecnológico, titulado: *Claves de una Justicia resiliente en plena transformación* (IP Sonia Calaza), del Ministerio de Ciencia e Innovación, con REF PID2024-155197OB-I00 y en la «Red de investigación»: *Alianzas estratégicas de la Justicia: Educación, Igualdad e Inclusividad* (RED2024-153961-T), coord. Sonia Calaza, Programa Estatal de Transferencia y Colaboración del Ministerio de Ciencia, Innovación y Universidades, Plan Estatal de Investigación Científica, Técnica y de Innovación 2024-2027.

Resumen

La Fiscalía de Barcelona (España) ha propuesto implementar una «calculadora electrónica de multirreincidencia» basada en inteligencia artificial para agilizar los procesos penales relacionados con hurtos reiterados, un problema que satura los juzgados catalanes y genera una percepción de impunidad. Esta herramienta verificaría rápidamente si un detenido cumple los requisitos para ser imputado por un delito menos grave, pero su naturaleza técnica y su impacto han generado dudas entre operadores jurídicos. Mientras algunos expertos cuestionan si se trata de verdadera inteligencia artificial o mera automatización, otros advierten sobre riesgos para los derechos fundamentales, la privacidad y la transparencia. Este ensayo analiza en profundidad las implicaciones jurídicas, éticas y prácticas de la propuesta, comparándola con experiencias internacionales y el marco normativo europeo, y aboga por un enfoque equilibrado que combine innovación tecnológica con garantías procesales.

Palabras clave

Multirreincidencia, Reglamento de Inteligencia Artificial, proceso penal, protección de datos, garantías procesales

Abstract

The Barcelona Public Prosecutor's Office (Spain) has proposed the implementation of an «electronic persistent-offending calculator» based on artificial intelligence to streamline criminal proceedings related to persistent theft offenses, an issue that burdens Catalan courts and contributes to a perceived sense of impunity. This tool would rapidly verify whether a detainee meets the legal requirements to be charged with a less serious offense; however, its technical nature and potential impact have raised concerns among legal practitioners. While some experts question whether the system constitutes genuine artificial intelligence or merely automated processing, others warn of potential risks to fundamental rights, privacy, and transparency. This essay provides an in-depth analysis of the legal, ethical, and practical implications of the proposal, comparing it with international experiences and the European regulatory framework, and ultimately advocates for a balanced approach that combines technological innovation with robust procedural safeguards.

Keywords

Persistent offending, AI Act, criminal procedure, data protection, procedural guarantees

I. Introducción a la propuesta de la Fiscalía de Barcelona y su relevancia en el contexto de la multirreincidencia

La multirreincidencia en delitos de hurto constituye uno de los principales desafíos estructurales del sistema judicial catalán, caracterizado por una saturación crónica de los juzgados que genera retrasos significativos en la resolución de casos y fomenta una creciente percepción social de impunidad. La reiteración delictiva, especialmente en los delitos patrimoniales de menor entidad, no solo erosiona la confianza ciudadana en la Administración de Justicia, sino que pone en evidencia la dificultad del sistema penal para ofrecer respuestas ágiles, proporcionadas y coherentes a fenómenos de criminalidad cotidiana y reiterada. Como ha señalado Landin Carrasco (1974) en su estudio criminológico sobre la multirreincidencia, este fenómeno no es nuevo, pero sus dimensiones actuales desbordan las capacidades de respuesta del sistema judicial tradicional.

En este contexto, la fiscal jefe de Barcelona, Neus Pujal, ha planteado una propuesta innovadora destinada a afrontar este problema mediante la implementación de una «calculadora electrónica de multirreincidencia» basada en inteligencia artificial. Esta herramienta, actualmente en estudio por el Departamento de Justicia de la Generalitat de Cataluña, tiene como objetivo verificar de manera automatizada y rápida si un detenido cumple los requisitos legales para ser imputado por un delito menos grave con penas privativas de libertad, conforme a lo dispuesto en el artículo 235 del Código Penal. El sistema pretende reducir la carga administrativa de los fiscales, homogeneizar los criterios de calificación jurídica y agilizar los procesos penales, todo ello con el propósito de combatir la reiteración delictiva y reforzar la respuesta institucional frente a la delincuencia reincidente.

La propuesta de la Fiscalía de Barcelona se inscribe en un contexto global de creciente interés por el uso de la inteligencia artificial en la administración de justicia, pero también de profunda preocupación por los riesgos que esta tecnología plantea para los derechos fundamentales. En distintas jurisdicciones del mundo se han ensayado modelos de automatización judicial con resultados dispares. China, Singapur o el Reino Unido han experimentado con herramientas de inteligencia artificial orientadas a la optimización de procesos, la gestión de expedientes o la predicción de riesgos de reincidencia; mientras que otras jurisdicciones, como la de Kerala (India), han adoptado una postura restrictiva, prohibiendo su uso en la adopción de decisiones judiciales por considerarlo incompatible con las garantías procesales.

La reforma penal en materia de multirreincidencia ha sido objeto de continuos debates doctrinales y jurisprudenciales. Como señala Magro Servet (2026), la evolución legislativa de la multirreincidencia en el Código Penal español refleja la tensión entre la necesidad de ofrecer respuestas punitivas efectivas y el respeto a los principios de proporcionalidad y *non bis in idem*. La Ley Orgánica 9/2022, de 28 de julio, modificó el artículo 234.2 del Código Penal para endurecer las penas en casos de multirreincidencia, pero su aplicación práctica ha sido desigual y ha generado numerosas controversias judiciales, como han documentado Maraver Gómez (2023) y Souto García (2019).

En el Reglamento de Inteligencia Artificial de 2024 se clasifica expresamente el uso de la inteligencia artificial en el ámbito judicial como de alto riesgo a tenor de su artículo 6.2 y el Anexo III apartado 6, imponiendo exigencias de transparencia, supervisión humana y responsabilidad institucional. Como explican Jiménez López (2024) y Simón Castellano (2024), este marco normativo europeo marca el límite entre la legítima modernización del sistema judicial y la potencial vulneración de derechos fundamentales. El concepto jurídico de inteligencia artificial en el Reglamento, analizado por Villalobos Portales (2025), es lo suficientemente amplio como para abarcar herramientas como la calculadora propuesta, siempre que incorporen técnicas de aprendizaje automático o inferencia autónoma.

En este escenario, la calculadora de multirreincidencia plantea un dilema central: ¿es posible aprovechar las ventajas de la inteligencia artificial para acelerar los procesos penales sin comprometer las garantías procesales, la independencia judicial y los derechos de los acusados? El interrogante trasciende el plano tecnológico y se adentra en el núcleo mismo del Estado de Derecho, donde la eficiencia no puede erigirse como valor supremo si amenaza los principios de legalidad, proporcionalidad y tutela judicial efectiva.

La relevancia de esta iniciativa no reside únicamente en su carácter experimental, sino en su capacidad para abrir un debate estructural sobre los límites del uso de la inteligencia artificial en el proceso penal. Este artículo analiza sus dimensiones jurídicas, éticas y prácticas, explorando tanto su potencial transformador como los desafíos que plantea en un marco normativo particularmente exigente. Inspirado en un enfoque analítico y claro, el texto busca desentrañar las complejidades técnicas, institucionales y de legitimidad democrática que acompañan a esta propuesta, contribuyendo al debate sobre el futuro de la justicia penal en la era algorítmica.

II. Contexto del problema de la multirreincidencia en el sistema judicial catalán

1. La saturación de los juzgados y la percepción de impunidad

El sistema judicial catalán enfrenta un problema estructural de saturación derivado del elevado volumen de casos relacionados con la multirreincidencia, particularmente en delitos de hurto. Según datos recientes del Consejo General del Poder Judicial, los juzgados catalanes gestionan anualmente miles de casos de hurtos reiterados, muchos de los cuales involucran a los mismos delincuentes que reinciden de manera sistemática. Esta acumulación de casos genera retrasos significativos en la resolución de los procesos penales, lo que contribuye a una percepción social de impunidad que erosiona la confianza de los ciudadanos en la justicia. La demora en la respuesta judicial no solo afecta a las víctimas, sino que también refuerza la reiteración delictiva, ya que los multirreincidentes perciben que las consecuencias de sus actos son mínimas o tardías.

La multirreincidencia ha sido estudiada en profundidad por la doctrina penal española. González-Cuéllar García (1976) realizó un estudio jurisprudencial

pionero sobre la multirreincidencia en el Código Penal español, identificando ya entonces las dificultades interpretativas que seguirían planteándose décadas después. García Arán (2007) analizó la reforma penal de 2003 en materia de multirreincidencia, señalando que el endurecimiento de las penas no siempre iba acompañado de los necesarios mecanismos procesales para garantizar su aplicación efectiva y proporcionada. La propia García Arán (2011) se preguntaba si es posible aplicar la agravación por multirreincidencia de forma coherente, una cuestión que sigue abierta tras las sucesivas reformas legislativas.

Rodríguez Mourullo (1972) abordó los aspectos críticos de la elevación de pena en casos de multirreincidencia, advirtiendo ya entonces sobre el riesgo de que la acumulación de condenas previas pudiera llevar a penas desproporcionadas en relación con la gravedad del último delito cometido. Esta preocupación ha sido reiterada por autores posteriores, como Aguado López (2008), quien en su monografía sobre la multirreincidencia y la conversión de faltas en delito planteó los problemas constitucionales y las alternativas político-criminales a este modelo punitivo.

Jordi Nieva, catedrático de Derecho Procesal de la Universidad de Barcelona, subraya que la lentitud en la respuesta judicial es una anomalía que perpetúa el problema de la multirreincidencia. En este contexto, la propuesta de la Fiscalía de Barcelona busca abordar esta situación mediante la automatización de la verificación de antecedentes penales, permitiendo a los fiscales identificar rápidamente a los reincidentes que cumplen los requisitos para enfrentar penas más severas. Sin embargo, la saturación de los juzgados no puede resolverse únicamente con herramientas tecnológicas, ya que requiere un enfoque integral que incluya el aumento de recursos humanos y materiales, como la creación de nuevos juzgados acordada para 2026 por el Consejo General del Poder Judicial, el Estado, la Generalitat y el Ayuntamiento de Barcelona.

2. Antecedentes de propuestas tecnológicas para abordar la multirreincidencia

La calculadora de multirreincidencia no es la primera iniciativa tecnológica destinada a abordar problemas judiciales en España. En el ámbito de la violencia de género, el sistema VioGén utiliza algoritmos para evaluar el riesgo de las víctimas y determinar las medidas de protección necesarias. Sin embargo, este sistema ha sido objeto de críticas por su falta de transparencia, su limitada eficacia en la predicción de riesgos y su impacto en los derechos de los justiciables. De manera similar, el algoritmo Veripol, empleado por la Policía Nacional para detectar denuncias falsas, fue retirado en marzo de 2025 tras demostrarse su falta de validez en procedimientos judiciales. Estos antecedentes ilustran los riesgos de implementar herramientas tecnológicas sin una validación rigurosa, un diseño técnico robusto y un marco normativo claro que garantice el respeto a los principios del debido proceso.

La experiencia con VioGén y Veripol pone de manifiesto la necesidad de abordar la integración de la inteligencia artificial con extrema cautela, especialmente en el ámbito penal, donde las decisiones pueden implicar la privación de

libertad. Cualquier herramienta automatizada aplicada al ámbito penal debe superar los errores observados en experiencias anteriores, garantizando que su diseño técnico y su implementación cumplan con los estándares éticos y jurídicos exigidos por la legislación nacional y europea.

Estos precedentes muestran que la introducción de inteligencia artificial en la justicia penal debe realizarse bajo una estricta validación técnica y jurídica. La eficiencia no puede anteponerse a las garantías procesales: una herramienta automatizada solo será legítima si respeta los derechos fundamentales, permite la trazabilidad de sus resultados y se somete a mecanismos efectivos de control público y judicial.

La experiencia acumulada con sistemas como VioGén y Veripol evidencia que la implementación tecnológica sin transparencia, supervisión ni auditoría puede generar nuevas disfunciones en lugar de resolver las ya existentes. En particular, la opacidad algorítmica, la insuficiente explicabilidad de los resultados y la posible reproducción de sesgos estructurales constituyen riesgos especialmente sensibles en el ámbito penal, donde las decisiones afectan directamente a derechos fundamentales y, potencialmente, a la libertad personal.

En consecuencia, cualquier incorporación de sistemas automatizados al proceso penal exige no solo una infraestructura tecnológica adecuada, sino también protocolos de validación rigurosos, auditorías periódicas y mecanismos claros de responsabilidad institucional. La automatización judicial no puede concebirse como una solución autosuficiente a los problemas estructurales de la Administración de Justicia, sino únicamente como un instrumento auxiliar sometido permanentemente al control humano y jurisdiccional.

3. La calculadora electrónica de multirreincidencia: origen y objetivos

La calculadora electrónica de multirreincidencia propuesta por la Fiscalía de Barcelona tiene como objetivo principal agilizar la identificación de delincuentes reincidentes que cumplan los requisitos establecidos en el artículo 235 del Código Penal español, que regula el delito de hurto en supuestos de multirreincidencia. Según este precepto, un detenido puede ser imputado por un delito menos grave si tiene al menos tres condenas firmes por hurtos que superen un total de 400 euros y sean computables. La herramienta automatizaría este proceso de verificación, permitiendo a los fiscales adoptar decisiones más rápidas y homogéneas sobre la imputación, lo que podría reducir la carga administrativa y mejorar la coherencia en la aplicación de la normativa penal.

El Departamento de Justicia de la Generalitat ha confirmado que está trabajando en el desarrollo de esta herramienta, aunque no ha especificado un plazo para su implementación. Sin embargo, la falta de concreción sobre su diseño técnico, los datos que utilizará y los mecanismos de supervisión ha generado reticencias entre los operadores jurídicos, quienes exigen garantías de que la herramienta respetará los derechos fundamentales y cumplirá con los principios de transparencia y auditabilidad.

En este sentido, la calculadora de multirreincidencia se diferencia formalmente de aquellos sistemas porque no pretende emitir valoraciones predictivas o subjetivas, sino verificar un dato jurídico objetivo —la existencia de condenas firmes previas—. Sin embargo, incluso esta función aparentemente neutra plantea interrogantes importantes. Su fiabilidad depende de la calidad y actualización de las bases de datos judiciales: cualquier error o falta de interoperabilidad puede conducir a imputaciones indebidas o a omisiones graves. Además, el grado de autonomía del sistema determinará su régimen jurídico: si incorpora mecanismos de aprendizaje automático, inferencia autónoma o adaptación tras su despliegue, será considerado un sistema de alto riesgo conforme al *AI Act*, y estará sometido a exigencias reforzadas de transparencia, documentación técnica, trazabilidad y supervisión humana.

Por otra parte, su utilización en el ámbito penal exige respetar escrupulosamente las garantías constitucionales: el derecho de defensa, la presunción de inocencia y el principio de intervención mínima del Derecho penal. Aun tratándose de un proceso de verificación aparentemente técnico, sus resultados pueden influir en decisiones que afectan directamente a la libertad de las personas. De ahí que sea indispensable garantizar la revisión humana sustantiva y la explicabilidad del proceso de decisión, evitando la automatización acrítica de juicios penales.

En conclusión, la calculadora electrónica de multirreincidencia constituye una propuesta orientada a mejorar la eficiencia procesal en la gestión de los supuestos de delincuencia reiterada. No obstante, su eventual implementación exige garantizar la fiabilidad técnica del sistema, la protección efectiva de los derechos fundamentales y la existencia de mecanismos adecuados de supervisión humana y control institucional. En consecuencia, su legitimidad y utilidad práctica dependerán de que la automatización permanezca subordinada a las garantías propias del Estado de Derecho y a un modelo de supervisión humana efectiva.

III. Análisis jurídico de la propuesta de inteligencia artificial

1. Naturaleza técnica de la calculadora: ¿automatización o inteligencia artificial?

Un aspecto central del debate en torno a la calculadora de multirreincidencia es su naturaleza técnica. Expertos como Jordi Nieva y Mireia Artigot Golobardes, profesora de Derecho Civil de la Universidad Pompeu Fabra, cuestionan si la herramienta propuesta constituye verdadera inteligencia artificial o si se limita a una automatización basada en una base de datos estructurada. Si la calculadora se reduce a un sistema de cálculo sistematizado que verifica condenas previas y aplica criterios predefinidos, no sería estrictamente una herramienta de inteligencia artificial, sino un *software* de gestión similar a los utilizados en otros ámbitos administrativos, como el programario de Microsoft Office. Esta distinción es crucial, ya que el uso de inteligencia artificial plantea

desafíos éticos y jurídicos más complejos, como la opacidad de los algoritmos y la posibilidad de que evolucionen autónomamente.

El concepto jurídico de inteligencia artificial en el Reglamento de Inteligencia Artificial de la Unión Europea, analizado en profundidad por Villalobos Portales (2025), es determinante para esta calificación. Según este autor, la definición de sistema de inteligencia artificial abarca aquellos sistemas diseñados para operar con distintos niveles de autonomía y que pueden, tras su despliegue, adaptarse generando patrones de comportamiento inferidos a partir de los datos. Si la calculadora se limita a ejecutar reglas predefinidas sin capacidad de aprendizaje o adaptación, quedaría fuera de esta definición y se regiría por las normas generales de automatización administrativa.

Nieva argumenta que, si la calculadora es un sistema automatizado, su implementación no debería generar un debate significativo, ya que herramientas similares han sido utilizadas durante años en la gestión judicial sin cuestionar los principios fundamentales del derecho. Sin embargo, si la herramienta incorpora algoritmos de aprendizaje automático que infieren resultados o evolucionan con el tiempo, su uso en el ámbito penal requeriría una regulación estricta para garantizar el respeto al debido proceso. Artigot advierte que la percepción de la inteligencia artificial como una solución mágica puede llevar a decisiones precipitadas, ignorando la necesidad de protocolos claros, sistemas de validación robustos y una supervisión humana efectiva.

La distinción entre automatización e inteligencia artificial tiene, además, implicaciones prácticas relevantes en términos de responsabilidad. Como señalan Ruiz Forns y Nicolás (2024), el nuevo Reglamento Europeo de Inteligencia Artificial establece un régimen de responsabilidad diferenciado según el nivel de riesgo del sistema. Los sistemas de alto riesgo, entre los que se incluyen los utilizados en la administración de justicia, están sujetos a obligaciones reforzadas de transparencia, documentación y supervisión humana. Por el contrario, los meros sistemas automatizados quedan fuera del ámbito de aplicación del Reglamento, aunque deben cumplir con los principios generales del Derecho administrativo y procesal.

2. Implicaciones para el debido proceso y los derechos fundamentales

La integración de la inteligencia artificial en el proceso penal plantea riesgos significativos para el debido proceso, consagrado en el artículo 24 de la Constitución Española, que garantiza el derecho a un juicio justo, la presunción de inocencia y el derecho de defensa. En el ámbito penal, donde las decisiones pueden implicar la privación de libertad, la cautela debe ser máxima. La calculadora de multirreincidencia, al automatizar la verificación de antecedentes penales, podría agilizar los procedimientos, pero también introduce el riesgo de errores algorítmicos que afecten los derechos fundamentales de los acusados. Por ejemplo, un falso positivo en la identificación de condenas previas podría resultar en una imputación injusta, mientras que un falso negativo podría permitir la liberación indebida de un reincidente.

Artigot subraya que la simple validación de los resultados generados por la inteligencia artificial no constituye una decisión humana plena, ya que los jueces o fiscales podrían verse influenciados por las recomendaciones algorítmicas, comprometiendo su autonomía. Además, la falta de transparencia en el funcionamiento de los algoritmos podría dificultar la impugnación de las decisiones, vulnerando el derecho de los acusados a cuestionar las pruebas en su contra. La experiencia con herramientas como VioGén y Veripol demuestra que los sistemas automatizados pueden generar resultados imprecisos o sesgados, lo que en el contexto de la multirreincidencia podría tener consecuencias graves para la libertad de las personas.

La jurisprudencia del Tribunal Supremo ha abordado en numerosas ocasiones las dificultades interpretativas de la multirreincidencia. La Sentencia 94/2025, de 6 de febrero, analizada en la Revista Aranzadi Doctrinal (2025), abordó el incremento de pena por multirreincidencia en delito leve de hurto, estableciendo criterios para el cómputo de condenas previas. La Sentencia 684/2019, de 3 de febrero de 2020, comentada por Larráyoz Sola (2020), declaró la inaplicabilidad del subtipo agravado por multirreincidencia en los delitos leves de estafa por considerarlo desproporcionado. La Sentencia 481/2017, de 28 de junio, analizada por Salinero Alonso (2018), abordó la tensión entre el principio de proporcionalidad, el *non bis in idem* y la multirreincidencia en el delito de hurto.

Estos pronunciamientos judiciales evidencian que la aplicación de la multirreincidencia no es una operación mecánica, sino que requiere interpretación jurídica y ponderación de circunstancias. La Sentencia 329/2025, de 9 de abril, comentada por Ortega Calderón (2025), abordó la aparente revisión del cómputo de cancelación de antecedentes penales, mostrando la complejidad de determinar cuándo una condena previa es computable a efectos de multirreincidencia. Una herramienta automatizada que no capture estas sutilezas jurídicas podría producir resultados contrarios a la jurisprudencia del Tribunal Supremo.

El derecho de defensa, consagrado en el artículo 24 de la Constitución, exige que el acusado pueda conocer y cuestionar todos los elementos que fundamentan la acusación. Si la calculadora de multirreincidencia opera como una "caja negra" cuyos criterios de decisión no son accesibles ni comprensibles para la defensa, se vulneraría este derecho fundamental. La defensa debe poder impugnar no solo el resultado de la herramienta, sino también el proceso lógico que ha llevado a ese resultado, incluyendo los datos utilizados, los algoritmos aplicados y los márgenes de error del sistema.

3. Requisitos de transparencia, auditabilidad y supervisión humana

Los operadores jurídicos, incluidos Jutgesses i Jutges per a la Democràcia y el Colegio de la Abogacía de Barcelona, han enfatizado la importancia de que la calculadora de multirreincidencia cumpla con estándares rigurosos de transparencia, auditabilidad y supervisión humana. La transparencia requiere que el algoritmo sea comprensible y que sus procesos de cálculo puedan ser explicados a los justiciables, permitiendo la impugnación de los resultados. La auditabilidad

implica que el sistema esté sujeto a revisiones periódicas para detectar posibles errores o sesgos, garantizando su fiabilidad. Finalmente, la supervisión humana asegura que las decisiones finales sean tomadas por jueces o fiscales, y no por la herramienta en sí, preservando la autonomía judicial.

Estas exigencias se alinean con los principios establecidos en la legislación europea sobre inteligencia artificial, que clasifica su uso en la justicia como de «alto riesgo» y exige controles estrictos. La política del Tribunal Superior de Kerala, que prohíbe el uso de inteligencia artificial en decisiones judiciales, ofrece un precedente relevante, destacando la importancia de mantener la supervisión humana como pilar del proceso judicial. En el caso de la calculadora de multirreincidencia, los operadores jurídicos demandan información detallada sobre el *corpus* de datos utilizado, los métodos de entrenamiento del algoritmo, los mecanismos para garantizar la calidad de los resultados y la responsabilidad de los operadores en caso de errores.

El Reglamento de Inteligencia Artificial, en su artículo 6.2 y Anexo III, apartado 6, clasifica como sistemas de alto riesgo aquellos utilizados para «evaluar el riesgo de reincidencia de una persona física o su comportamiento delictivo» o para «establecer el perfil de una persona física en el contexto de la investigación, detección o enjuiciamiento de infracciones penales». La calculadora de multirreincidencia, al verificar la concurrencia de condenas previas, podría caer dentro de esta categoría, lo que activaría las obligaciones reforzadas del Reglamento. Jiménez López (2024) analiza en profundidad la interacción entre el Reglamento de Inteligencia Artificial y el Reglamento General de Protección de Datos, señalando que ambos instrumentos deben aplicarse de manera coordinada para garantizar una protección efectiva de los derechos fundamentales.

La Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD, 2018) establece, en su artículo 9, un régimen especial para el tratamiento de datos penales. La calculadora de multirreincidencia, al acceder a antecedentes penales y condenas firmes, trataría datos especialmente protegidos, lo que exige medidas de seguridad reforzadas y, en determinados casos, la realización de una evaluación de impacto relativa a la protección de datos. El artículo 35 del Reglamento General de Protección de Datos exige esta evaluación cuando el tratamiento, por su naturaleza, alcance o finalidades, pueda implicar un alto riesgo para los derechos y libertades de las personas físicas.

La extraterritorialidad del Reglamento de Inteligencia Artificial, analizada por Ortega Giménez (2024), implica que cualquier herramienta de inteligencia artificial utilizada en el ámbito judicial español debe cumplir con los estándares europeos, independientemente de que su desarrollo o implementación sea realizada por una autoridad nacional. Esto refuerza la necesidad de que la calculadora de multirreincidencia se diseñe e implemente conforme a los principios de transparencia, auditabilidad y supervisión humana exigidos por el Reglamento.

La supervisión humana efectiva, exigida por el artículo 14 del Reglamento de Inteligencia Artificial, no puede ser meramente formal. No basta con que un fiscal o juez valide mecánicamente el resultado de la herramienta; es necesario que tenga la capacidad de comprender, cuestionar y, en su caso, apartarse de

la recomendación algorítmica. Esto implica que los operadores jurídicos deben recibir formación específica sobre el funcionamiento de la herramienta, sus limitaciones y sus posibles sesgos. Como señala Simón Castellano (2024), la supervisión humana debe ser «sustantiva», es decir, debe implicar un verdadero control sobre el proceso de decisión, no una mera rúbrica formal.

IV. Retos éticos y prácticos de la integración de la inteligencia artificial

1. Riesgos de sesgos algorítmicos y protección de datos

Uno de los principales riesgos asociados con la calculadora de multirreincidencia es la posibilidad de que los algoritmos perpetúen sesgos presentes en los datos de entrenamiento. Por ejemplo, si la base de datos utilizada refleja desigualdades históricas en la aplicación de la justicia penal —como una mayor incidencia de condenas contra ciertos grupos sociales por razones socioeconómicas o culturales—, el algoritmo podría generar resultados discriminatorios, contraviniendo el principio de igualdad ante la ley consagrado en el artículo 14 de la Constitución española. Artigot subraya la importancia de conocer el *corpus* de datos con el que trabajaría la herramienta, así como los métodos de entrenamiento empleados, para garantizar que no se vulneren los derechos fundamentales.

Sánchez Benítez (2020) ha analizado la relación entre aporofobia (rechazo o discriminación hacia las personas pobres) y derecho penal, señalando que el delito de hurto y la circunstancia agravante de multirreincidencia afectan desproporcionadamente a los sectores más vulnerables de la población. Una herramienta automatizada que no incorpore mecanismos correctores podría amplificar esta discriminación estructural, convirtiendo la eficiencia algorítmica en un instrumento de injusticia social. El autor propone incorporar cláusulas de equidad algorítmica que verifiquen periódicamente la ausencia de sesgos discriminatorios.

La protección de datos personales es otra preocupación central. La calculadora de multirreincidencia requeriría acceder a información sensible, como antecedentes penales, datos personales y detalles de los procesos judiciales, lo que plantea riesgos de filtraciones, uso indebido o acceso no autorizado. El Reglamento General de Protección de Datos de la Unión Europea establece requisitos estrictos para el tratamiento de datos en contextos judiciales, exigiendo que se adopten medidas técnicas y organizativas para garantizar la seguridad y confidencialidad. La falta de claridad sobre cómo se gestionarán estos datos en la propuesta de la Fiscalía genera incertidumbre sobre su viabilidad y su conformidad con la normativa europea.

Mercader Uguina y Velasco Pardo (2025) analizan el impacto en lo laboral del Reglamento de Inteligencia Artificial, pero sus reflexiones sobre la necesidad de evaluaciones de impacto y la garantía de la intervención humana son extensibles al ámbito judicial. López Nieto (2025) examina la figura de los operadores de inteligencia artificial según el Reglamento europeo, distinguiendo entre proveedores, usuarios e importadores, y asignando a cada uno responsabilidades

específicas. En el caso de la calculadora de multirreincidencia, la Generalitat de Cataluña actuaría como usuario de un sistema de inteligencia artificial, lo que implica obligaciones de supervisión, formación y rendición de cuentas.

2. Limitaciones en la reducción de los tiempos judiciales

Aunque la calculadora de multirreincidencia busca agilizar los procesos penales, su impacto real en la reducción de los tiempos judiciales podría ser limitado. Jutgesses i Jutges per a la Democràcia y Jordi Nieva coinciden en que la herramienta no abordará las causas estructurales de la saturación de los juzgados, como la falta de jueces, fiscales y personal administrativo. La deliberación judicial, que implica analizar pruebas, interpretar normas y ponderar circunstancias específicas, seguirá siendo necesaria, independientemente de la automatización de ciertas tareas. En este sentido, la calculadora podría optimizar la fase inicial de verificación de antecedentes, pero no resolverá los cuellos de botella en las etapas posteriores del proceso penal, como la celebración de juicios o la emisión de sentencias.

Además, la experiencia con herramientas como VioGén y Veripol sugiere que los sistemas automatizados no siempre cumplen con las expectativas de eficiencia. La necesidad de supervisión humana y auditorías rigurosas podría introducir nuevas capas de complejidad administrativa, contrarrestando los beneficios de la automatización. Para que la calculadora sea efectiva, debe integrarse en una estrategia más amplia que incluya la creación de nuevos juzgados y la mejora de la infraestructura judicial, como ha acordado el Consejo General del Poder Judicial para 2026.

Mercader Uguina (2025) señala, en su análisis del Reglamento de Inteligencia Artificial, que «frecuentemos el futuro» con cautela, sin dejarnos deslumbrar por las promesas tecnológicas. Esta reflexión es aplicable al ámbito judicial: la eficiencia algorítmica no debe convertirse en un fin en sí mismo, sino en un medio para mejorar la calidad de la justicia. Si la automatización introduce nuevos riesgos (errores, sesgos, opacidad) sin ofrecer beneficios claros en términos de celeridad o acierto, su implementación será contraproducente.

3. La necesidad de una estrategia integral más allá de la tecnología

La presión por encontrar soluciones rápidas a la multirreincidencia no debe llevar a la adopción de medidas superficiales que prioricen la tecnología sobre una estrategia integral. Artigot advierte contra la percepción de la inteligencia artificial como una solución mágica, subrayando que su implementación debe ir acompañada de protocolos claros, sistemas de validación robustos y una evaluación de impacto. Sin una planificación adecuada, la calculadora de multirreincidencia podría convertirse en una medida estética que no aborde las causas profundas del problema, como la falta de recursos judiciales, las dinámicas sociales que fomentan la reincidencia o la necesidad de políticas de prevención del delito.

Una estrategia integral requeriría combinar la tecnología con medidas estructurales, como el aumento del número de jueces, la mejora de la formación

de los operadores jurídicos y la implementación de programas de reinserción social para reducir la reincidencia. Nieva destaca que la rapidez en los procesos judiciales es clave para reducir la percepción de impunidad, pero esto no puede lograrse únicamente con herramientas tecnológicas. La calculadora debe ser vista como un complemento, no como un sustituto, de las reformas necesarias para modernizar el sistema judicial catalán y garantizar una justicia efectiva y equitativa.

Las actitudes hacia los delincuentes multirreincidentes en España han sido estudiadas por Serrano Maíllo (2020), quien señala que la opinión pública tiende a favorecer respuestas punitivas severas, pero también valora las políticas de reinserción cuando se demuestra su eficacia. Esta dualidad debe ser tenida en cuenta en el diseño de cualquier herramienta tecnológica: la calculadora no debe ser percibida como un instrumento de mero endurecimiento penal, sino como un mecanismo para aplicar la ley de manera más eficiente y equitativa.

V. Perspectivas comparadas: experiencias internacionales y regulación europea

1. Uso de inteligencia artificial en otros sistemas judiciales

La propuesta de la Fiscalía de Barcelona se inscribe en un contexto global donde varias jurisdicciones han experimentado con la inteligencia artificial en el ámbito judicial, ofreciendo lecciones valiosas para Cataluña. En China, los tribunales utilizan sistemas de análisis predictivo para recomendar sentencias en casos civiles y administrativos, lo que ha permitido agilizar la resolución de disputas en un sistema con un volumen masivo de casos. Sin embargo, estas herramientas han sido criticadas por su opacidad y su falta de supervisión humana efectiva, lo que plantea riesgos para los derechos fundamentales. En Singapur, el poder judicial ha implementado inteligencia artificial para redactar borradores de sentencias y gestionar casos, siempre bajo estricta supervisión humana, logrando un equilibrio entre eficiencia y responsabilidad.

En el Reino Unido, herramientas como CaseCrunch analizan grandes volúmenes de datos jurídicos para proporcionar recomendaciones a los jueces, demostrando el potencial de la tecnología para mejorar la consistencia y la rapidez en la toma de decisiones. Estas jurisdicciones han establecido regulaciones claras para garantizar la transparencia y la rendición de cuentas, lo que contrasta con la postura restrictiva del Tribunal Superior de Kerala, que prohíbe el uso de inteligencia artificial en decisiones judiciales. La propuesta de Barcelona se sitúa en un punto intermedio, buscando aprovechar la tecnología para agilizar procesos sin delegar decisiones judiciales, pero enfrenta desafíos similares en términos de transparencia, auditabilidad y protección de derechos.

El impacto del Reglamento de Inteligencia Artificial en las entidades locales ha sido analizado por Moro Cordero (2024), quien señala que la implementación de sistemas de inteligencia artificial en la administración pública debe ir acompañada de evaluaciones de impacto, formación de los empleados públicos y mecanismos de transparencia. Estas exigencias son extensibles a la Administración

de Justicia, que en Cataluña es competencia de la Generalitat. La calculadora de multirreincidencia, si se implementa, deberá cumplir con estos estándares.

2. El marco normativo europeo sobre los sistemas de inteligencia artificial de alto riesgo

La legislación europea sobre inteligencia artificial, adoptada en 2024, clasifica su uso en el ámbito judicial como de «alto riesgo», lo que implica restricciones significativas y requisitos estrictos de supervisión. Según el Reglamento de Inteligencia Artificial de la Unión Europea, las herramientas utilizadas en la justicia deben cumplir con estándares de transparencia, auditabilidad y protección de datos, garantizando que las decisiones finales sean tomadas por humanos y no por algoritmos. Este marco normativo plantea un desafío para la calculadora de multirreincidencia, ya que su implementación requerirá demostrar que no compromete los derechos fundamentales ni introduce sesgos algorítmicos.

La experiencia con VioGén y Veripol ilustra los riesgos de no cumplir con estos estándares. Ambos sistemas fueron criticados por su falta de transparencia y su impacto en los derechos de los justiciables, lo que llevó a su cuestionamiento o retirada. La calculadora de multirreincidencia deberá superar estas limitaciones, incorporando mecanismos de validación y auditoría que garanticen su conformidad con la normativa europea. Esto incluye la publicación de informes detallados sobre el funcionamiento del algoritmo, la verificación de los datos de entrenamiento y la implementación de sistemas de supervisión que permitan detectar y corregir errores en tiempo real.

Los sistemas de inteligencia artificial prohibidos o inaceptables en el Reglamento son analizados por Simón Castellano (2024), quien enumera las prácticas de inteligencia artificial que se consideran contrarias a los valores europeos, como la manipulación subliminal, la explotación de vulnerabilidades o la puntuación social. La calculadora de multirreincidencia no incurre en estas prohibiciones, pero su clasificación como sistema de alto riesgo implica obligaciones significativas en términos de gobernanza de datos, documentación técnica y supervisión humana. El resto de los operadores de inteligencia artificial, como proveedores e importadores, tienen responsabilidades específicas que deben ser asignadas claramente en el caso de la calculadora.

3. Lecciones aplicables al caso de Cataluña

Las lecciones de la jurisprudencia del Tribunal Supremo sobre multirreincidencia, resumidas en los trabajos de De Vicente Martínez (2021, 2022), Capita Remezal (2018), Liñán Lafuente (2020), Alabau Pereiro (2025) y Rodríguez Centeno (2023), evidencian que la aplicación de esta figura no es una operación mecánica. La calculadora debe ser capaz de reflejar los matices jurisprudenciales, como la imposibilidad de computar condenas por delitos leves para la agravación del hurto multirreincidente (Larráyoz Sola, 2019), o la inaplicabilidad del subtipo agravado por multirreincidencia en los delitos leves de estafa por

desproporción (Larráyoz Sola, 2020). Sin esta capacidad de capturar las sutilezas jurídicas, la herramienta generaría más problemas que soluciones.

Las reflexiones de la doctrina sobre la necesidad de una reforma penal coherente en materia de multirreincidencia (Magro Servet, 2026; De Vicente Martínez, 2022; Vázquez Berdugo, 2016) apuntan a que el problema no es solo de eficiencia procesal, sino también de política criminal. Una calculadora electrónica, por muy precisa que sea, no resolverá las disfunciones de fondo si el sistema penal sigue sin ofrecer respuestas proporcionadas y efectivas a la delincuencia reiterada. La tecnología debe ir acompañada de una reflexión política y jurídica sobre el sentido y los límites de la agravación por multirreincidencia.

Las experiencias internacionales y el marco normativo europeo ofrecen lecciones valiosas para la implementación de la calculadora de multirreincidencia en Cataluña. En primer lugar, la transparencia es esencial para garantizar que los justiciables comprendan cómo se toman las decisiones que afectan sus derechos, permitiendo la impugnación de los resultados algorítmicos. En segundo lugar, la supervisión humana debe ser un pilar central, evitando que los jueces o fiscales se limiten a validar resultados generados por la herramienta. Finalmente, la protección de datos y la prevención de sesgos requieren un diseño técnico robusto, con auditorías periódicas y un *corpus* de datos libre de desigualdades históricas.

El precedente de Kerala destaca la importancia de establecer límites claros para el uso de la inteligencia artificial, asegurando que no se deleguen decisiones judiciales a sistemas automatizados. Por otro lado, las experiencias de Singapur y el Reino Unido demuestran que es posible integrar la tecnología de manera responsable, siempre que se adopten regulaciones estrictas y se priorice la supervisión humana. Cataluña puede aprender de estos modelos, adoptando un enfoque híbrido que combine la eficiencia de la inteligencia artificial con las garantías procesales necesarias para proteger los derechos fundamentales y mantener la confianza pública en el sistema judicial.

La evaluación de impacto algorítmico, similar a la evaluación de impacto relativa a la protección de datos prevista en el artículo 35 del Reglamento General de Protección de Datos, debería ser obligatoria antes de la implementación de la calculadora. Esta evaluación debería analizar no solo la fiabilidad técnica del sistema, sino también sus efectos sobre los derechos fundamentales, la equidad procesal y la confianza pública en la justicia. Precisamente, tendría que ser realizada por un equipo independiente, con participación de expertos en derecho penal, protección de datos, ética algorítmica y representantes de la sociedad civil.

VI. Reflexiones finales sobre el equilibrio entre innovación tecnológica y garantías procesales

La propuesta de la Fiscalía de Barcelona para implementar una calculadora electrónica de multirreincidencia representa un intento audaz de abordar uno de los mayores desafíos del sistema judicial catalán: la saturación de los juzgados y la percepción de impunidad derivada de la multirreincidencia. Sin embargo, su

éxito dependerá de la capacidad de superar los retos técnicos, éticos y jurídicos que plantea la integración de la inteligencia artificial en el ámbito penal. La herramienta debe diseñarse con transparencia, auditabilidad y supervisión humana como principios rectores, garantizando que no comprometa los derechos fundamentales ni perpetúe sesgos algorítmicos. Al mismo tiempo, debe integrarse en una estrategia integral que aborde las causas estructurales de la saturación de los juzgados, como la falta de recursos humanos y materiales.

La experiencia internacional y el marco normativo europeo ofrecen un marco de referencia para implementar la calculadora de manera responsable, equilibrando la innovación tecnológica con las garantías procesales. La cautela expresada por los operadores jurídicos no debe interpretarse como una oposición a la tecnología, sino como una demanda de rigor en su aplicación. La calculadora de multirreincidencia tiene el potencial de transformar la lucha contra la reincidencia, pero solo si se implementa con un compromiso firme con los principios de justicia, equidad y transparencia. En última instancia, el desafío radica en aprovechar las ventajas de la inteligencia artificial sin sacrificar los valores fundamentales que sustentan el sistema judicial, asegurando que la tecnología sirva a la justicia y no al revés.

La reforma del artículo 234.2 del Código Penal por la Ley Orgánica 9/2022, de 28 de julio, analizada por Maraver Gómez (2023) y Souto García (2019), ha pretendido clarificar el régimen de la multirreincidencia en los delitos leves de hurto, pero su aplicación práctica sigue planteando dudas. La STS 94/2025, de 6 de febrero (Revista Aranzadi Doctrinal, 2025), ha vuelto a insistir en la necesidad de interpretar restrictivamente esta agravación, para evitar penas desproporcionadas. La calculadora debe ser capaz de reflejar esta orientación jurisprudencial, no limitándose a un cómputo mecánico de condenas previas.

En conclusión, la calculadora electrónica de multirreincidencia es una propuesta que merece ser estudiada con rigor, pero también con prudencia. Su implementación no puede ser precipitada ni deslumbrarse por las promesas de eficiencia tecnológica. Debe ir precedida de una evaluación de impacto exhaustiva, de un diseño técnico transparente y auditado, y de un marco normativo que garantice la supervisión humana y la protección de los derechos fundamentales. Solo así podrá convertirse en una herramienta útil para modernizar la justicia penal catalana, sin traicionar los principios que la sustentan.

Referencias bibliográficas

Aguado López, S. (2008). *La multirreincidencia y la conversión de faltas en delito: problemas constitucionales y alternativas político-criminales*. Iustel.

Alabau Pereiro, P. (2025). La circunstancia de multirreincidencia en el delito de hurto: una cuestión no resuelta. *Revista jurídica Universidad Autónoma de Madrid*, (52), 43-68.

Capita Remezal, M. (2018). La agravante de multirreincidencia en el delito de hurto. Una propuesta de regulación. *La ley penal: revista de derecho penal, procesal y penitenciario*, (132).

De Vicente Martínez, R. (2021). El hurto agravado por la multirreincidencia y la pena de prohibición de acudir al lugar donde se cometió el delito. *Revista de derecho y proceso penal*, (62), 129-154.

De Vicente Martínez, R. (2022). El final de una errónea interpretación jurisprudencial: la reforma del artículo 234.2 del Código Penal. *Diario La Ley*, (10128).

García Arán, M. (2007). La multirreincidencia en la reforma penal de 2003. *Revista galega de seguridade pública: REGASP*, (9), 127-136.

García Arán, M. (2011). ¿Es posible aplicar la agravación por multirreincidencia de forma coherente? En H. Hormazábal Malarée (Coord.), *Estudios de derecho penal: en memoria del Prof. Juan José Bustos Ramírez* (pp. 93-110). Universidad Externado de Colombia.

González-Cuéllar García, A. (1976). *La multirreincidencia en el código penal español (estudio jurisprudencial).* (Tesis doctoral). Universidad Autónoma de Madrid.

Jiménez López, J. (2024). El Reglamento de inteligencia artificial y el Reglamento general de protección de datos. En L. Cotino Hueso y P. Simón Castellano (Dirs.), *Tratado sobre el reglamento de inteligencia artificial de la Unión Europea* (pp. 149-180). Tirant lo Blanch.

Landin Carrasco, A. (1974). *Estudio criminológico sobre la multirreincidencia.* (Tesis doctoral). Universidad Complutense de Madrid.

Larráyoz Sola, I. (2019). No deben computarse las condenas por delitos leves para la aplicación del tipo agravado de hurto multirreincidente: STS 579/2018, de 21 de noviembre. *Revista Aranzadi Doctrinal*, (2).

Larráyoz Sola, I. (2020). No cabe aplicar el subtipo agravado por multirreincidencia en los delitos leves de estafa por considerarlo desproporcionado: STS 684/2019, de 3 febrero 2020. *Revista Aranzadi Doctrinal*, (4).

Liñán Lafuente, A. (2020). La agravante de multirreincidencia en los delitos leves contra el patrimonio. *Revista Sistema Penal Crítico*, (1), 253-265.

López Nieto, Y. (2025). Los operadores de inteligencia artificial según el Reglamento europeo de inteligencia artificial. En T. Bastarreche Bengoa, A. Egea de Haro y M. M. Serrano Pérez (Dirs.), *El Reglamento europeo de inteligencia artificial: un análisis crítico y multidisciplinar* (pp. 109-139). Universitas.

Magro Servet, V. (2026). Análisis de la reforma penal en materia de multirreincidencia que viene. *Diario La Ley*, (10886).

Maraver Gómez, M. (2023). La regulación de la multirreincidencia en los delitos de hurto tras la reforma producida por la Ley Orgánica 9/2022, de 28 de julio. *Revista electrónica de ciencia penal y criminología*, (25).

Mercader Uguina, J. R. (2025). El reglamento de inteligencia artificial, frecuentemos el futuro. En M. E. Casas Baamonde (Coord.), *Los Briefs de la Asociación Española de Derecho del Trabajo y de la Seguridad Social: Las claves de 2024* (pp. 185-190). Tirant lo Blanch.

Mercader Uguina, J. R. y Velasco Pardo, B. (2025). El impacto en lo laboral del reglamento de Inteligencia Artificial. *Revista Derecho Social y Empresa*, (23), 106-133.

Moro Cordero, M. A. (2024). El impacto del Reglamento de Inteligencia Artificial en las entidades locales. *Consultor de los ayuntamientos y de los juzgados*, (extra 3).

Ortega Calderón, J. L. (2025). La aparente revisión del cómputo de cancelación de antecedentes penales: STS 329/25 de 9 de abril. *Diario La Ley*, (10734).

Ortega Giménez, A. (2024). La extraterritorialidad del nuevo Reglamento europeo de Inteligencia Artificial. *Revista jurídica: Región de Murcia*, (54), 121-145.

Revista Aranzadi Doctrinal. (2025). Incremento de pena por multirreincidencia en delito leve de hurto: Tribunal Supremo, Sala de lo Penal, Sentencia 94/2025, 6 Feb. Rec. 5454/2022. *Revista Aranzadi Doctrinal*, (4).

Rodríguez Centeno, R. (2023). Penalidad de la multirreincidencia en los delitos leves de hurto. Modificación 234 Código Penal. ¿Hacia una nueva y efectiva aplicación? *Diario La Ley*, (10246).

Rodríguez Mourullo, G. (1972). Aspectos críticos de la elevación de pena en casos de multirreincidencia. *Anuario de derecho penal y ciencias penales*, *25*(2), 289-304.

Ruiz Forns, A. y Nicolás, A. (2024). Nuevo Reglamento Europeo de Inteligencia Artificial. *Diario La Ley*, (10491).

Salinero Alonso, C. (2018). A vuelta con el principio de proporcionalidad, de "non bis in idem" y la multirreincidencia en el delito de hurto: análisis de la STS 481/2017, de 28 de junio. En P. M. de la Cuesta Aguado, L. R. Ruiz Rodríguez, M. Acale Sánchez, E. Hava García, M. J. Rodríguez Mesa, G. González Agudelo, I. Meini Méndez y J. M. Ríos Corbacho (Coords.), *Liber amicorum: estudios jurídicos en homenaje al profesor doctor Juan Ma. Terradillos Basoco* (pp. 921-934). Aranzadi.

Sánchez Benítez, C. (2020). Aporofobia y derecho penal: el delito de hurto y la circunstancia agravante de multirreincidencia. *Revista Sistema Penal Crítico*, (1), 225-240.

Serrano Maíllo, A. (2020). Actitudes hacia los delincuentes multirreincidentes en España: Un enfoque explicativo. *Indret: Revista para el Análisis del Derecho*, (2).

Simón Castellano, P. (2024). El resto de los sistemas de inteligencia artificial prohibidos o inaceptables en el reglamento de inteligencia artificial. En L. Cotino Hueso y P. Simón Castellano (Dirs.), *Tratado sobre el reglamento de inteligencia artificial de la Unión Europea* (pp. 275-288). Tirant lo Blanch.

Souto García, E. M. (2019). La multirreincidencia en los delitos de hurto tras la L.O. 1/2015, de 30 de marzo: su regulación y aplicación práctica. *La ley penal: revista de derecho penal, procesal y penitenciario*, (141).

Vázquez Berdugo, I. (2016). Delitos de robo y hurto: multirreincidencia: evolución legislativa: especial análisis de la reforma introducida por L.O 1/15, de 30 de marzo. *Revista del Ministerio Fiscal*, (1), 104-155.

Villalobos Portales, J. (2025). Concepto jurídico de inteligencia artificial en el Reglamento de inteligencia artificial de la Unión Europea. En M. Pastrana Espárraga y E. Olmedo Peralta (Dirs.), *El Derecho de la Competencia ante las plataformas digitales* (pp. 247-284). Tirant lo Blanch.

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 156-172

USO EFICIENTE DE LA INTELIGENCIA ARTIFICIAL EN LA EVALUACIÓN PROBATORIA DEL PROCESO ADMINISTRATIVO SANCIONADOR: EL CASO PERUANO[1]

EFFICIENT USE OF ARTIFICIAL INTELLIGENCE IN THE EVIDENTIARY EVALUATION OF THE ADMINISTRATIVE SANCTIONING PROCESS: THE PERUVIAN CASE

Carmen Milagros Velarde Koechlin[2]

Universidad de Salamanca

1 Se utilizó Chat GPT (Open AI) como herramienta de apoyo en la sistematización de información y estructuración del análisis. No obstante, la interpretación jurídica, selección de fuentes y construcción argumentativa corresponden exclusivamente a la autora.

2 Doctoranda en Derecho Privado de la Universidad de Salamanca, España. Docente de la Unidad de Virtualización Académica de la Universidad de San Martín de Porres, Perú. Abogada. Máster en Derecho Empresarial y Magíster en Gerencia Social. Magíster Artis en Gobernabilidad y Procesos Electorales. Es jefa nacional del Registro Nacional de Identificación y Estado Civil del Perú (RENIEC). ORCID N° 0000-0002-2668-7774.

Resumen

El presente artículo analiza el uso de la inteligencia artificial (IA) en la evaluación de los medios probatorios en el procedimiento administrativo sancionador en el Perú, a partir del estudio de la Resolución Final n.º 083-2025/CC3 del Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual (INDECOPI), para determinar las implicancias jurídicas de la utilización de herramientas de IA en el análisis de la evidencia digital, especialmente en la detección de llamadas *spam*. En ese sentido, se evalúa si dicho uso respeta los principios del debido proceso, como el derecho a la supervisión humana, la transparencia algorítmica y el derecho de defensa dentro del nuevo proceso administrativo algorítmico. La metodología empleada es de carácter cualitativo, basada en el análisis normativo, doctrinal y jurisprudencial del caso, así como en la revisión de literatura especializada sobre inteligencia artificial y prueba digital. Como resultado, se evidencia que la incorporación de IA mejora la eficiencia en la gestión probatoria, pero genera riesgos asociados a la opacidad algorítmica, sesgos y falta de explicabilidad. Se propone la implementación de estándares mínimos como la supervisión humana efectiva, la transparencia del sistema, la responsabilidad demostrada y la posibilidad de contradicción por parte del administrado, a fin de garantizar la legitimidad de las decisiones administrativas en un entorno de gobernanza algorítmica.

Palabras clave

Inteligencia artificial, responsabilidad demostrada, procedimiento administrativo sancionador, supervisión humana, transparencia algorítmica

Abstract

This article analyzes the use of artificial intelligence (AI) in evidentiary evaluation within the administrative sanctioning procedure in Peru, based on a study of INDECOPI's Final Resolution No. 083-2025/CC3. The aim is to determine the legal implications of using AI tools in the analysis of digital evidence, particularly in the detection of spam calls. It assesses whether this use respects due process principles, such as the right to human oversight, algorithmic transparency, and the right to defense within the new algorithmic administrative process. The methodology employed is qualitative, based on a normative, doctrinal, and jurisprudential analysis of the case, as well as a review of specialized literature on artificial intelligence and digital evidence. The results show that the incorporation of AI improves the efficiency of evidence management but generates risks associated with algorithmic opacity, bias, and a lack of explainability. The implementation of minimum standards is proposed, such as effective human supervision, system transparency, technical validation, and the possibility of challenge by the citizen, in order to guarantee the legitimacy of administrative decisions in an algorithmic governance environment.

Keywords

Artificial intelligence, proven responsibility, administrative sanctioning procedure, human supervision, algorithmic transparency

I. Introducción

Desde la aparición de la inteligencia artificial (IA) y sus modelos más desarrollados —IA *machine learning*, IA generativa e IA agentiva— las ciencias jurídicas han optado por el uso de estos mecanismos como coadyuvador en el trabajo legal. Sin duda, la gran capacidad de almacenaje de datos y de evaluación de información para presentar propuestas de contenidos masivos y soluciones en el menor tiempo posible, han sido base para incorporar la IA en los diferentes ámbitos donde el Derecho se desarrolla y operativiza (Sánchez Acevedo, 2022). Eficiencia y eficacia; es decir, mayor trabajo y mejores resultados con menor presupuesto e impactos superiores o sobresalientes con menores tiempos.

Los organismos administrativos del Estado y tribunales de justicia han comenzado a incorporar modelos de IA para que ésta contribuya a facilitar su labor, generalmente, a través de sistemas que —basados en información previa de resoluciones administrativas o judiciales anteriores, precedentes o jurisprudencia— proponen ante un caso determinado, el proyecto de resolución decisoria cuyo contenido, por supuesto, deberá pasar luego por la revisión humana antes de convertirse en el fallo final (Aguilar Cabrera, 2024).

Este pase a la revisión humana se ha transformado en un principio ético para el uso de la IA y también en un nuevo derecho: el derecho a la supervisión humana de la IA, en línea con la necesidad de garantizar la protección de derechos fundamentales frente a decisiones automatizadas (Sánchez Acevedo, 2022). Ello, como respuesta a los casos donde se han presentado demandas, proyectos de ley u otros documentos jurídicos hechos en IA y sin revisión humana, que han incorporado información inexacta, inexistente o errada, por el fenómeno de alucinación. Es decir, la IA crea contenido propio e inexistente para no dejar vacíos. Por eso, todo producto final elaborado por un modelo de inteligencia artificial debe pasar obligadamente por una supervisión humana.

Vemos entonces cómo la IA participa en niveles jurídicos como el nivel procedimental: donde coopera en la elaboración de documentos; y el nivel normativo que abarca la generación de leyes y otras normas que regulan el uso de la IA, incluyéndose aquí las decisiones administrativas y judiciales que analizan la IA y los nuevos derechos digitales que concibe. Pero existe otro nivel de uso de la IA en el campo jurídico y es el ámbito probatorio, cuando ésta contribuye al análisis de los medios de prueba presentados, lo cual exige protocolos claros para garantizar la validez y fiabilidad de la evidencia digital (Aguilar Cabrera, 2024), nivel que ha sido aplicado en el Perú a través de la Resolución Final n.º 083-2025/CC3, emitida por la Comisión de Protección al Consumidor del Instituto Nacional de Defensa de la Competencia y de la Propiedad Intelectual (Indecopi).

En el presente artículo, se analiza dicho caso y los debates a los que conllevó tal uso para la evaluación de medios probatorios, destacando en nuestro

análisis la necesidad de demostrar que el sistema informático de IA se encuentra operativo, sin errores o sesgos que pudieran afectar el debido proceso o derecho de defensa en concordancia con la exigencia de transparencia, explicabilidad y control de los algoritmos en la prueba digital; y que el derecho a la supervisión humana va más allá de revisar el contenido brindado por la IA; es la posibilidad de mantener la credibilidad de las decisiones de la autoridad administrativa y la fiabilidad del administrado, en las herramientas tecnológicas, cada vez más avanzadas; es decir, lograr la legitimidad de una gobernanza algorítmica.

II. Ideas clave: el procedimiento administrativo algorítmico y la prueba digital

La transición de las administraciones públicas hacia entornos digitales no debe entenderse como una mera digitalización de archivos, sino como la consolidación del procedimiento administrativo algorítmico. En este nuevo paradigma, la eficiencia y eficacia son los motores de cambio, pero estos principios deben armonizarse con las garantías del debido proceso en la gestión de la evidencia digital. El desarrollo teórico de este apartado se sustenta en los siguientes ejes fundamentales:

1. Transparencia y explicabilidad algorítmica frente a la opacidad

La utilización de herramientas de inteligencia artificial en la función administrativa introduce el riesgo de la opacidad tecnológica o problemáticas que pueden conllevar a sesgos algorítmicos o a discriminaciones, dificultando el entendimiento del funcionamiento interno del modelo. Frente a ello, Cerrillo i Martínez (2021) sostiene que la transparencia debe ser funcional y proactiva, permitiendo que el administrado no sólo acceda al código, sino que comprenda la lógica de la decisión. Esta visión complementa lo expuesto por Sánchez Acevedo (2022), quien advierte que el uso de la IA en el sector público encuentra su límite infranqueable en el respeto a los derechos fundamentales. Así, la transparencia algorítmica se configura como una extensión del derecho de acceso a la información pública.

2. La Prueba asistida y la responsabilidad demostrada

La incorporación de sistemas de *speech to text* y procesamiento de lenguaje natural (PLN) transforma la prueba tradicional en una prueba asistida algorítmicamente. Aguilar Cabrera (2024) destaca que este nivel probatorio exige protocolos claros que garanticen la validez y fiabilidad de la evidencia digital. Este aspecto se enlaza con el concepto de responsabilidad demostrada propuesto hace ya varios años por Nelson Remolina (2022); la autoridad no sólo debe aplicar la tecnología, sino demostrar que el sistema ha sido debidamente calibrado para evitar sesgos o alucinaciones que afecten la presunción de inocencia.

3. El derecho a la supervisión humana y la reserva de humanidad

Un pilar crítico de este marco es el derecho a la supervisión humana, el cual garantiza que las decisiones punitivas no sean plenamente automatizadas. Siguiendo a Gamero Casado (2021), es fundamental distinguir entre el simple acompañamiento humano y una supervisión jurídica real; la supervisión requiere mecanismos determinados, fases y controles que evidencien una evaluación completa. En sintonía, Piñar Mañas (2008) ya había subrayado que la reserva de humanidad en el proceso sancionador es un mandato que asegura que la IA actúe como un coadyuvante y no como un sustituto del juicio humano.

Desde este marco conceptual introductorio, es significativo presentar un caso suscitado en Perú, específicamente la Resolución final n.º 083-2025/CC3 de Indecopi. Así, se muestra el caso señalado en tanto que representa un hito, pues pone a prueba la capacidad de la administración para equilibrar la eficiencia tecnológica con las garantías propuestas por la doctrina citada.

III. El caso peruano: uso de la IA para la evaluación eficiente de medios probatorios en la detección de llamadas *spam*. Resolución final n.º 083-2025/CC3, del Indecopi

El estudio del presente caso resulta emblemático para el derecho administrativo peruano, al representar uno de los primeros hitos donde el uso de inteligencia artificial es el eje central de una sanción por llamadas masivas no autorizadas o *spam*. A fin de examinar la actuación de la autoridad y las garantías del administrado, esta sección se organiza en tres etapas. La primera detalla los antecedentes y hechos que dieron origen al procedimiento, centrándose en la recolección de las más de 560,000 pruebas digitales; la segunda, analiza la implementación técnica y el uso de los modelos Whisper y RStudio para el procesamiento de datos; finalmente, la tercera etapa examina los argumentos de las partes, contrastando la defensa de la empresa frente a la motivación expuesta por la autoridad en la resolución de sanción.

1. Antecedentes y hechos: el procesamiento de grabaciones masivas y el uso de la IA

El Instituto Nacional de Defensa de la Competencia y de la Propiedad Intelectual (Indecopi), es la entidad pública peruana encargada de valorar los casos de derecho de la competencia, propiedad intelectual y protección al consumidor, cuidando que las reglas del mercado no sean asimétricas ni abusivas y cumplan con las normas legales. En el caso de protección al consumidor, uno de los temas que se le ha encargado al Indecopi es la revisión de asuntos relacionados al recibimiento de correos electrónicos no deseados o publicidad no deseada, lo que además incluye a las llamadas telefónicas no solicitadas que ofrecen productos y servicios, es decir, comunicaciones *spam*. Con ello, es vital para el Indecopi verificar primero que la empresa que hace el contacto cuente con el consentimiento expreso —léase, aceptación— del consumidor para recibir tal comunicación.

Para ello, Indecopi actúa a pedido de parte cuando recibe alguna denuncia, pero también de oficio, como parte de su labor supervisora, para lo cual solicita a las empresas, los audios de las llamadas telefónicas comerciales, las que son escuchadas para su evaluación y confirmación de que se trata de una llamada publicitaria y que se cuenta con la aceptación del consumidor para recibirlas.

Dentro de este panorama, la Resolución Final n.º 083-2025/CC3 del Indecopi constituye un caso paradigmático sobre el uso de inteligencia artificial en la actividad administrativa sancionadora en el Perú. El caso se inicia a partir de un procedimiento administrativo sancionador de oficio contra el Banco BBVA Perú (BBVA, en adelante), por la presunta realización de comunicaciones comerciales sin consentimiento previo de los consumidores, infracción prevista en el artículo 58 del Código de Protección y Defensa del Consumidor.

Para ello, el Indecopi solicitó al BBVA que le enviara las grabaciones de las comunicaciones telefónicas realizadas a los consumidores de acuerdo con los centros de contacto que hubiese contratado o que hubiese realizado directamente. Las comunicaciones debían corresponder a los períodos del 15 al 31 de enero de 2024; del 01 al 15 de marzo de 2024; y, del 15 de abril al 30 de abril del 2024, (Indecopi, 2025c). En el Perú, las entidades que utilizan canales telefónicos para realizar comunicaciones comerciales deben mantener tales grabaciones como medios probatorios de la aceptación de servicios o productos determinados, es decir, la grabación es el medio de contratación telefónica.

El BBVA remitió al Indecopi un total de 569.401 grabaciones de llamadas. Este número es importante porque muestra una cantidad de minutos telefónicos más que elevado que deben —en un procedimiento normal— ser escuchados por humanos para verificar o descartar que se traten de llamadas comerciales. Este es un primer filtro. Es decir, no se da por hecho que el universo total de llamadas entregadas sea publicitario, sino que debe revisarse una por una para verificar que lo sea y que el consumidor, previamente, haya otorgado su consentimiento de recibir la comunicación.

2. Procesamiento de datos y uso de instrumentos de la IA

Así, para el procesamiento de datos, el Indecopi consideró necesario el uso de la IA y la tecnología en general para ayudar en la revisión de las grabaciones telefónicas, utilizando modelos de reconocimiento de voz y procesamiento de lenguaje natural. Destacó el uso de Whisper (modelo *large-v3-turbo* en su versión optimizada *Faster-Whisper 1.0.3*) para la transcripción automatizada de audios y también el uso de lenguajes de programación como Python y R (RStudio) para la clasificación y muestreo estadístico de la información. El Indecopi, para filtrar las comunicaciones, utilizó expresiones publicitarias como: «préstamo disponible, líneas de crédito, tarjeta de crédito, campaña, renovación, membresía, entre otras» (Indecopi, 2025b), detectándose con las herramientas digitales, un total de «330,934 audios con contenido publicitario» (Indecopi, 2025b).

El segundo filtro utilizado por el Indecopi está relacionado a principios estadísticos. Por ello, de este nuevo universo de 330,934 audios, «se seleccionó una muestra estadísticamente representativa correspondiente a los tres períodos

fiscalizados (…) con un tamaño de 385 elementos por período (1,155 comunicaciones en total)» (Indecopi, 2025b). Igualmente, 1155 audios implican una gran cantidad de minutos a ser escuchados y evaluados por lo que se mantuvo la apuesta por la utilización de la IA.

Gracias al uso de la IA Whisper y demás herramientas tecnológicas, se logró revisar las 1155 grabaciones, encontrando responsabilidad de BBVA por no haber solicitado el consentimiento del consumidor para ofertar telefónicamente sus productos y servicios, lo que conllevó a la imposición de una multa equivalente a poco más de US$ 400.000 (cuatrocientos mil dólares americanos) o € 384,000 (trescientos ochenta y cuatro mil euros).

Hasta aquí el caso parece sencillo y refleja la utilización de la IA con eficiencia: más grabaciones analizadas en menor tiempo posible y con un alto índice de certeza sobre su contenido vulnerador de los derechos del consumidor.

Sin embargo, es importante analizarlo más profundamente, desde los sustentos de defensa del administrado hasta los fundamentos de la autoridad para respaldar el uso, control y resultados obtenidos con apoyo de la IA, destacando que, desde ambas partes, no sólo se ha argumentado jurídicamente, sino también de forma técnica, lo que en adelante deberá ser una exigencia para los abogados y las abogadas.

Ante el uso de las tecnologías como medios probatorios o como apoyo en el análisis de pruebas deberá argumentarse también técnicamente; ello para encontrar sus falencias o garantizar que no tengan vacío alguno que tienda a un posible mal funcionamiento que afecte los resultados y, con ello, el derecho a un debido proceso y el derecho de defensa. De igual modo, la autoridad decisora, deberá adoptar todas las acciones necesarias que conlleven a la credibilidad y confianza en la utilización y resultados que arrojen los sistemas informáticos, los aplicativos y, sobre todo, los sistemas de IA por ellos empleados. Así, será imprescindible que todos los operadores jurídicos manejen conceptos y normas relacionadas al derecho digital y principios y procesos técnicos y operativos como sistemas de gestión que garanticen la auditoría y supervisión de las herramientas tecnológicas.

3. Argumentos de las partes. El debate entre el administrado y la autoridad

Durante el caso peruano estudiado, el administrado cuestionó el uso de la IA como parte del proceso evaluativo sosteniendo que (Indecopi, 2025b):

a) No debió excluirse el universo completo de llamadas pues sesga la representatividad estadística y el análisis integral con relación a la posible infracción.
b) El Indecopi no habría sustentado el motivo por el cual eligió determinados términos para considerar que los audios serían publicitarios; más aún, sólo habría sustentado que se usaron herramientas informáticas.
c) No se habría considerado los lineamientos emitidos por el propio Indecopi sobre el uso de la IA como herramienta complementaria sin sustituir el juicio humano, evitando además «decisiones automatizadas sin supervisión

humana» (Indecopi, 2025b). En este último punto sostuvo que la propia Organización para la Cooperación y el Desarrollo Económico (OCDE) destaca el uso de la IA en la administración pública bajo supervisión humana y dando las facilidades para la impugnación de las decisiones que se basen en herramientas tecnológicas.

d) Sobre Whisper, concretamente estableció que ésta no será una herramienta fiable, ya que existen investigaciones que han demostrado problemas en la precisión de las transcripciones con un error de 1% al generar alucinaciones, es decir, contenidos que no responderían a lo comprendido en los audios originales. El administrado, incluso, adjuntó un cuadro comparando lo expuesto por la IA versus lo descrito realmente en el audio demostrando transcripciones erradas.

e) Destacó el administrado que las demás herramientas informáticas como RStudio también presentaban limitaciones en la construcción de los procesos de clasificación, por lo que requería una supervisión humana.

f) Insistió que, a pesar de que había habido acompañamiento humano, el resultado final había arrojado errores, por lo que recomendó hacer una «revisión manual exhaustiva y detallada de las conversaciones transcritas para evitar sanciones en casos donde sí hubo consentimiento válido» (Indecopi, 2025b).

Los cuestionamientos del administrado son coincidentes con la doctrina comparada que observa que el uso de algoritmos en procesos jurídicos puede generar riesgos de opacidad, sesgos y afectación a derechos fundamentales si no existen mecanismos adecuados de control y transparencia (Sánchez Vilanova, 2022).

Sobre lo cuestionado, la autoridad administrativa refutó cada punto explicando lo siguiente (Indecopi, 2025b):

a) La autoridad sí evaluó el 100% del universo de las llamadas entregadas por el administrado. Dicha valoración se dio en el momento inicial, cuando todos los audios fueron verificados por los sistemas tecnológicos para ser disgregados en audios con contenido publicitario, por un lado, y audios que no incluyen contenido promocional, por otro. Aquí puede apreciarse cómo el uso de un modelo de IA contribuyó a revisar todos los medios probatorios (100% de las grabaciones de las llamadas telefónicas) para definir aquellos que serán considerados como válidos para el caso concreto. De no contarse con la IA, tal vez se hubiera tenido que revisar manualmente (humanamente) cada grabación, lo que afectaría posiblemente los plazos del proceso administrativo; o se habría utilizado la técnica estadística del muestreo, manejándose un número determinado de audios desde un inicio y eligiéndose cada uno al azar.

b) Con relación a la selección de términos para establecer el contenido publicitario, el Indecopi sostuvo que tales términos fueron notificados al BBVA como parte de los anexos a los informes enviados.

c) Sobre los cuestionamientos al uso de la IA, el Indecopi expresó que se había remitido al BBVA el código fuente que habían usado para la IA Whisper y para RStudio, así como las transcripciones obtenidas por los sistemas informáticos, como parte de los anexos a los informes remitidos al administrado.

Además, precisó que se había seguido los Lineamientos del Indecopi para el uso de la IA y las recomendaciones de la OCDE, aplicando la supervisión humana, tanto en la etapa de procesamiento como de validación de la data, pasando por dos equipos humanos durante el proceso sancionador: el equipo de la Dirección de Fiscalización e Instrucción y el equipo de la Secretaría Técnica de la Comisión de Protección al Consumidor.

d) Refiriéndose exclusivamente al sistema de IA Whisper, la autoridad aceptó que el sistema puede presentar algunas falencias que conlleven a discrepancias, sin embargo, para minimizar tales riesgos, hizo uso de la versión más actualizada de Whisper, la misma que garantiza mayor precisión en los análisis y adoptó diversos mecanismos de control, entre ellos, la supervisión humana. Mencionó también que la tasa de error era mínima y que tal desviación fue corregida, sin ocasionar perjuicio al administrado.

e) Sobre las demás herramientas informáticas, la autoridad resaltó nuevamente la supervisión humana realizada durante todo el proceso administrativos y en cada una de las fases de evaluación y análisis del contenido.

f) Destacó, finalmente, que la IA había complementado el proceso administrativo pero que éste, en todo momento, tuvo la intervención y supervisión humana en el análisis de la data (audio) y que las verificaciones aplicadas permitieron la corrección de cualquier resultado que pueda afectar contraria e indebidamente, al administrado, por lo que no correspondía nueva revisión manual.

Como puede apreciarse, la posición de la autoridad ha sido a favor del uso de los algoritmos defiendo su utilización y destacando que la misma es un complemento al análisis de los medios probatorios, pero que de ninguna manera se deja de lado la supervisión humana, la cual se aplica en diversas oportunidades y fases del proceso evaluativo y del proceso administrativo en general para garantizar un resultado sin errores. No obstante, los cuestionamientos del administrado nos permiten también reflexiones sobre la validación y legitimidad en el uso de la IA en procesos administrativos sancionadores, razonamientos que deben considerarse para la mejora de los nuevos derechos digitales a los que la IA conlleva.

Expuestos los antecedentes, la metodología técnica y los argumentos que sustentaron la decisión del Indecopi en la Resolución n.º 083-2025/CC3, queda en evidencia que no nos encontramos ante un caso administrativo ordinario, sino ante un punto de inflexión en la gestión de la evaluación probatoria. La resolución analizada no sólo resuelve una controversia sobre llamadas no autorizadas, sino que pone en evidencia la tensión latente entre la eficiencia que prometen las herramientas de inteligencia artificial y los límites que estas mismas tienen con relación al derecho administrativo, en particular a entidades sancionadoras. Por ello se pasa de mostrar el caso al análisis y discusión jurídica que, a la luz de la doctrina más reciente, evalúe si la praxis administrativa peruana está preparada para garantizar un debido proceso algorítmico frente a los desafíos de la opacidad, el margen de error y la necesaria reserva de humanidad.

IV. Los nuevos derechos digitales en el proceso administrativo algorítmico a ser considerados en el uso de la IA como herramienta para el análisis de medios probatorios

Cuando la administración pública decide aplicar la IA en sus actividades de fiscalización, de clasificación de información, de transcripción o de apoyo a la toma de decisiones, no es suficiente la búsqueda de eficiencia. Los procesos administrativos, sobre todo, los procesos sancionadores, implican una acción punitiva, una penalidad y, por tanto, un daño. Así, la eficiencia no es pura, se enfrenta a derechos y por ello, debe considerar la aplicación de estándares mínimos que cuiden la forma y fondo en la aplicación de la IA y otras herramientas tecnológicas que convierten al tradicional procedimiento administrativo en un procedimiento administrativo algorítmico. Por ejemplo, lograr la trazabilidad de los sistemas utilizados: ¿quién usó el sistema e ingresó a él?, ¿quién lo actualizó?, ¿quién colocó la información a procesar y quién recibió los resultados?, ¿a dónde fueron estos enviados? Cada paso debe estar trazado pues cualquier falla implica responsabilidad a la administración pública. Revisemos cómo la IA se hace parte del procedimiento administrativo en la generación y análisis de los medios probatorios.

1. La IA como herramienta probatoria en la administración pública

La incorporación de sistemas de inteligencia artificial (IA) en la función administrativa ha generado una transformación sustantiva en la producción, análisis y valoración de los medios probatorios. En el caso analizado, se evidencia el uso de herramientas de *speech to text* o transcripciones de audio a texto en procedimientos administrativos, lo que presenta un salto cualitativo desde la prueba tradicional hacia la prueba asistida algorítmicamente.

Con ello, puede considerarse a una evidencia algorítmica, como aquella que es generada o procesada a través del uso de sistemas informáticos y que pueden otorgar inferencias probabilísticas. Por tanto, al tener el poder de expresar probabilidades de conductas, de acciones y actividades de juzgamiento, se requiere de una revisión humana bastante aguda que certifique la veracidad de tales probabilidades y anule cualquier error o riesgo posible de una mala decisión basada en los resultados de la IA.

Estudios de casos como los juzgamientos europeos de SyRI donde según The Hague District Court (2020) y Sánchez Vilanova (2022) señalan que el Tribunal de Países Bajos declaró incompatible con el Convenio Europeo de Derechos Humanos el uso de un sistema de perfilado algorítmico por falta de transparencia y verificabilidad; o el caso SCHUFA donde el Tribunal de Justicia de la Unión Europea (2023) determinó que cuando una decisión se base en un algoritmo, debe activarse la prohibición de decisiones automatizadas del artículo 22 del Reglamento General de Protección de Datos, han demostrado que los sistemas de inteligencia artificial, sobre todo aquellos que generan su propio contenido, presentan problemáticas que pueden conllevar a discriminaciones o sesgos algorítmicos. Ello es conocido como la opacidad tecnológica (*black box*), ya que existe siempre una dificultad para entender el funcionamiento interno del modelo de

IA: Ninguna persona sin conocimientos de sistemas y sin conocimiento del código fuente, puede conocer o evaluar el algoritmo y cómo éste toma sus decisiones.

Al tratarse de sistemas informáticos que trabajan con base a probabilidades, dejan abierta una puerta de desconfianza sobre la fiabilidad y robustez del modelo de IA y del software o programa en sí. En el caso peruano, la Resolución Final n.º 083-2025/CC3 del Indecopi evidencia el uso de Whisper para la transcripción de audios en procedimientos sancionadores, reconociendo las fallas que tal programa presenta (errores) y destacando en varias partes de su decisión final, la intervención y control humano sobre los resultados automatizados (Indecopi, 2025b).

2. El debido procedimiento administrativo y la prueba algorítmica

Ante el panorama del uso de la IA en los procedimientos administrativos, se exige al Derecho reinterpretar el contenido del debido procedimiento administrativo. ¿Cómo se integra el nuevo proceso administrativo algorítmico? Tal vez resultó más sencillo en el pasado considerar el procedimiento electrónico, basado en digitalización de la documentación, uso de firma digital para autenticación del administrado y relacionarlo con el contenido; sin embargo, un procedimiento administrativo algorítmico incluye un sistema de asistencia que presenta pareceres, opiniones y posibles decisiones, basándose en situaciones que podrían ser, que podrían ocurrir. Por ello, es necesario que se considere mejoras en los principales derechos y deberes:

a) El derecho de defensa: Que debe permitir al administrado conocer cómo se generó un medio probatorio; qué sistemas de IA o herramientas informática se utilizaron para ello y la posibilidad de cuestionar la fiabilidad de los resultados. En el caso peruano, se ejemplifica con relación a los términos seleccionados para determinar si los audios contenían temas publicitarios. Indecopi sostuvo que los términos se entregaron en un anexo del informe notificado, pero no sustentó cómo los eligió y cuál fue el motivo por el cual los consideró relacionados a temas promocionales. Esto tiene que ver con el uso del *prompt* o instrucción dada a la IA. Sin duda, la búsqueda de información y los resultados serán más acertados de acuerdo con el tipo de *prompt* que se ingrese en la IA. De otro lado, tal información debió colocarse dentro del documento principal ya que todo uso de la IA debe ser lo más transparente posible.
b) El deber de motivación de las decisiones por parte de la autoridad administrativa: Al utilizarse modelos determinados de IA, será fundamental que las autoridades incluyan el uso de tales sistemas en sus resoluciones y decisiones. No basta con mencionar cuáles son dichos programas, sino explicar qué rol tuvieron dentro del procedimiento administrativos; qué medidas se adoptaron para contar con la seguridad de que los resultados obtenidos por dicha IA y otras herramientas informáticas, sean los correctos y tengan la evaluación jurídica y técnica precisa. El deber de motivación tendrá a su vez que explicar cuáles han sido los controles humanos utilizados y en qué fases del análisis y procedimiento se dieron. Si tal participación humana

resultó en un acompañamiento o conllevó a una auditoria previa y durante el procedimiento del sistema de IA, o si se aplicó una supervisión basada en protocolos previamente aprobados.

En el contexto peruano, si bien el Indecopi sostuvo en su Resolución que la IA no sustituye la decisión humana y que se actuó con base a lineamientos institucionales de uso de IA y con supervisión humana constante, no se explicó totalmente cómo se desarrollaron tales supervisiones o si existía un protocolo previo de supervisión. Por tanto, es necesario evaluar un estándar explícito sobre cómo integrar la evidencia algorítmica en la motivación administrativa a fin de lograr la comprensión por parte del administrado, de tal forma que pueda cuestionar la prueba algorítmica con base a sustentos jurídicos y técnicos.

Debe añadirse también que el uso de la IA en el caso peruano se basó en el establecimiento de lineamientos institucionales previos relacionados al uso de la IA en Indecopi, sentando los principios éticos y de supervisión y auditoría de los sistemas de IA a ser utilizados en la entidad. De este modo, dicha institución acreditó previamente el uso de cualquier IA mediante la Resolución n.º 000062-2025-GEG/Indecopi que aprobó los Lineamientos para el uso ético de la IA haciendo un reconocimiento formal del uso institucional de la IA y destacando la necesidad de generar trazabilidad, responsabilidad en su utilización y el deber del control humano. Además, previo a ello, Indecopi creó un equipo especial para el uso de la IA en la entidad, a través de la Resolución n.º 000151-2024-PRE/Indecopi.

c) El derecho a la supervisión humana, que garantiza todo control humano sobre el resultado final y que se basa en la necesidad de contar con sistemas informáticos – sobre todo, sistemas de IA – fiables y robustos. Para proteger este derecho habrá que preguntarse: ¿Cómo se valida la supervisión humana?, ¿existe algún protocolo que deban seguir los supervisores? Una situación es el acompañamiento humano y otra es la supervisión humana. La supervisión requiere un mecanismo determinado, fases, controles, formatos que evidencien que se cumplió el proceso que garantiza la evaluación completa y la minimización de errores, mientras que el acompañamiento humano es una intervención para obtener resultados del sistema informático, pero sin una revisión minuciosa ni la evaluación o contrastación de los posibles resultados.

De otro lado, el derecho a la supervisión humana existe porque concurre la desconfianza; el escepticismo previo a toda nueva tecnología, más aún, con respecto a la IA, ya que, si sus resultados deben ser revisados por humanos, implica ello que desde un inicio ya presentan margen de error que podría no ser detectado y afectaría a los administrados. Pensemos en el uso de una calculadora científica; ninguna persona dudaría que los resultados que brinde deban ser evaluados por presencia humana. Los números son los que son luego de las operaciones matemáticas respectivas. Igual ocurre con un cajero automático. Nadie pensaría que podría ofrecer más o menos dinero que el solicitado, si no, la cantidad exacta. La supervisión humana, si bien es un derecho, es también la aceptación de un funcionamiento no valido en su totalidad o de un mal funcionamiento de los sistemas de IA, tal vez debido a su novedad. Ello implica que

toda tecnología que sea utilizada en un proceso administrativo algorítmico debe garantizar la máxima efectividad y, para ello, debe estar debidamente calibrada y explicada a los administrados para reducir los riesgos. Podrá existir márgenes de error, pero estos deben ser mínimos, similares a los márgenes de error que podrían producir los humanos y que se atenúan basados en el criterio o parecer del decisor o juzgador que tiene tal libertad al momento de decidir.

Así, en el caso peruano, de no utilizarse la IA, los audios hubieran sido escuchados por personas, ingresando a la prueba el concepto de subjetividad, el considerar alguna palabra como una negativa de aceptar la publicidad. Resaltamos nuevamente, existe también el error humano o los posibles sesgos personales. No existe una evaluación 100% específica. Dado que el administrado busca su defensa y evitar la sanción, siempre estará cuestionando el uso tecnológico, por lo que debe existir un protocolo de calibración o protocolo de puesta a cero del sistema, de encendido y prueba, de certificación de contarse con el último y más actualizado código fuente o la constancia de que el sistema está listo para su uso de acuerdo a esa versión tecnológica y que tal versión, a pesar de no ser la más actualizada, no afecta los resultados del proceso administrativo.

d) El derecho de transparencia algorítmica. Implica la entrega del código fuente y de todo criterio de uso de la IA u otro programa informático que es utilizado en el proceso de análisis probatorio. En el caso peruano, se aprecia que la autoridad señaló haberlo entregado como parte de un anexo a un informe determinado. Consideramos que éste debe ser parte del documento principal y no de un anexo, a fin de que sea explícitamente visto y revisado por el administrado. No obstante, caben algunas preguntas: ¿tiene el administrado la capacidad para revisar los códigos fuente?, ¿conoce la IA?, ¿el plazo para evaluar tal IA es el adecuado?, ¿puede usar la IA para replicar el procedimiento de la autoridad? Todo ello debe complementar este derecho a la transparencia algorítmica, ya que es necesario que el administrado entienda qué sistemas informáticos se usan, por qué, para qué, qué producen, cómo lo favorecen o lo cuestionan.
e) El derecho al debido procedimiento algorítmico, lo que significa la posibilidad de contradicción de los resultados generados por la IA. Ello no necesariamente debe esperar a la siguiente instancia, sino que debe tenerse toda la información necesaria para cuestionar en cualquier parte del proceso administrativo, cualquier resultado generado por la IA.

Este derecho ha sido resaltado por la Corte Constitucional de Colombia (2025) en la Sentencia T-067/2025 como un derecho fundamental que reconoce la transparencia algorítmica como parte del derecho de acceso a la información pública, estableciendo que los ciudadanos pueden acceder a información sobre el funcionamiento de sistemas automatizados estatales.

En efecto, el uso de inteligencia artificial en la valoración probatoria exige el respeto a los principios del debido proceso, la presunción de inocencia y la posibilidad de contradicción, especialmente cuando se trata de evidencia digital generada o analizada mediante IA (Aguilar Cabrera, 2024). De otro lado, la doctrina señala que los algoritmos deben ser evaluados bajo criterios de fiabilidad

científica, transparencia y responsabilidad, a fin de garantizar su legitimidad en el proceso judicial o administrativo. Además, el uso de sistemas automatizados en la administración pública tiene como límite el respeto a los derechos fundamentales, especialmente cuando estos inciden en decisiones que afectan a los ciudadanos (Sánchez Acevedo, 2022)

El caso peruano nos muestra un sistema de regulación del procedimiento administrativo algorítmico para lograr eficiencia, utilizando la IA e incorporando constantemente la supervisión humana. No obstante, aunque Indecopi haya señalado en su Resolución que remitió al administrado el código fuente del sistema de IA utilizado, lo hizo dentro de los anexos de su Informe de investigación, lo que podría generar una opacidad relativa ya que éste debe ser explícitamente entregado al administrado. Por tanto, el procedimiento administrativo del caso peruano carece de transparencia algorítmica, pero también de criterios de auditoría y de reglas claras para el cuestionamiento probatorio.

El uso de IA en el caso peruano descrito mostraría que la administración peruana ha pasado la fase inicial, es decir, la fase experimental, ya que se ha animado a regular y motivar – aunque débilmente – en su resolución, los motivos del uso de la IA, la forma de utilización concreta y el control humano constante a los resultados de esta en cada fase de evaluación de los medios probatorios. Perú podría estar en una nueva fase de uso estructural de los sistemas algorítmicos en la producción de medios probatorios.

3. Responsabilidad demostrada y margen de error

En este punto es necesario mirar el principio de responsabilidad demostrable, concepto que, como sostiene Remolina (2022), obliga a las organizaciones que implementan IA no sólo a cumplir con la norma, sino además a estar en capacidad de demostrar dicho cumplimiento mediante evidencias claras de control y mitigación de riesgos. En el caso analizado, surge una tensión evidente entre la eficiencia administrativa y la precisión técnica: Indecopi admitió que el modelo Whisper presenta un margen de error cercano al 1%, lo que en un universo de 569.401 grabaciones se traduce en más de 5600 posibles transcripciones erróneas.

Bajo la óptica de la responsabilidad demostrable, no basta con que la autoridad alegue una alta tasa de éxito. Siguiendo la tesis de Remolina (2022), la administración debe probar que implementó una debida diligencia algorítmica antes de emitir la sanción. Esto implica que, ante el reconocimiento de un margen de error, la autoridad debió detallar qué medidas de auditoría técnica se aplicaron para asegurar que la empresa sancionada no fuera víctima de una alucinación del lenguaje o de una falsa identificación de términos publicitarios. La falta de un protocolo de verificación sobre ese 1% de error desplaza la carga de la prueba hacia el administrado, quien se ve obligado a desvirtuar una evidencia generada por una caja negra, lo cual debilita el estándar de probidad y transparencia que debe regir en un proceso sancionador en tiempos contemporáneos.

V. Conclusiones

El uso de la inteligencia artificial en los procedimientos administrativos no será nunca totalmente neutro, sus posibilidades de uso y su capacidad probabilística afectan la evaluación de los medios probatorios, haciendo eficiente el proceso, brindando información nueva o complicándolo a través de posibles alucinaciones.

La Resolución n.º 083-2025/CC3 constituye un precedente relevante en el derecho administrativo peruano y confirma que la administración pública ha trascendido la mera digitalización para pasar a una administración pública algorítmica pero también plantea desafíos críticos en torno a la auditabilidad y el control jurídico de las decisiones apoyadas en sistemas de IA. El caso confirma que el uso de inteligencia artificial como herramienta legítima en la actividad probatoria y de fiscalización, permite procesar acciones humanas en un tiempo eficaz, reduciendo el proceso complejo de procesamiento de volúmenes de información casi inacabables para el ojo humano. Sin embargo, se evidencia que se requiere que se integre la transparencia metodológica y algorítmica con miras a garantizar el debido procedimiento. Este caso evidencia la transición hacia una administración pública algorítmica.

El caso peruano evidencia un avance significativo en eficiencia, pero insuficiente con relación a los nuevos derechos digitales y al proceso administrativo algorítmico, por lo que se requiere la creación protocolos de supervisión humana de la IA. Así, en el caso revisado, hemos visto la necesidad de un protocolo detallado que explique cómo el personal humano validó las transcripciones del 1% de error detectado. Ello confirmar que no basta con una intervención humana típica o básica para observar resultados automáticos por lo que se concluye que el procedimiento sancionador peruano debe blindar una reserva de humanidad efectiva. En ese sentido, se propone que se debe proveer el nuevo derecho de la supervisión humana, entendida esta como una revisión crítica y con capacidad de correcciones respecto de los hallazgos generados por la inteligencia artificial.

El uso de la inteligencia artificial en la evaluación probatoria constituye una herramienta valiosa para mejorar la eficiencia del procedimiento administrativo sancionador, pero su implementación debe estar acompañada de garantías jurídicas que aseguren su transparencia, control y compatibilidad con los derechos fundamentales, en línea con los estándares doctrinales y normativos actuales (Aguilar Cabrera, 2024; Sánchez Acevedo, 2022).

Se confirma que la transparencia en el uso de algoritmos no se satisface con la entrega de códigos fuente en anexos técnicos, como fue el caso peruano mostrado, puesto que ello constituye únicamente una transparencia limitada y formalista. En ese sentido, se requiere una transparencia y explicabilidad, donde la autoridad traduzca el funcionamiento del algoritmo a un lenguaje jurídico comprensible para los administrados, además de garantizar el derecho a la contradicción técnica, permitiendo que el administrado pueda cuestionar y peritar los resultados del software sin barreras de opacidad. De lo contrario, el derecho de defensa se ve mermado. Con ello se plantea una necesidad de que futuras

resoluciones incluyan un resumen o mapa de la lógica algorítmica aplicada, de forma accesible y transparente.

En esa línea de ideas, se constata que el caso peruano revisado, es una muestra de que se requiere de una responsabilidad demostrada. Vimos que el margen de error del 1 % admitido por Indecopi no es una cifra trivial cuando se traduce en miles de posibles falsos positivos. Esta situación nos permite verificar que se requiere que la autoridad administrativa debe preocuparse por la fiabilidad de las herramientas tecnológicas que emplea y que, por ende, la administración debe adoptar un enfoque de responsabilidad demostrada que incluya auditorias técnicas previas con pruebas de estrés a los algoritmos de detección de *spam*. Al respecto, se propone que se prevea la supervisión humana reforzada, la trazabilidad de los sistemas utilizados, aplicabilidad de la decisión y la responsabilidad por los errores del sistema.

Solo mediante el respeto estricto a este los derechos y deberes identificados en este estudio de caso y junto a elementos conceptuales claves, la inteligencia artificial podrá consolidarse como una herramienta de eficiencia legítima, capaz de modernizar la gestión pública sin sacrificar las garantías fundamentales en el mundo actual.

Referencias bibliográficas

Aguilar Cabrera, D. A. (2024). La inteligencia artificial en la justicia: protocolos para la presentación y la valoración de prueba digital obtenida mediante IA. *Revista Oficial del Poder Judicial*, *16*(22), 475-497. https://doi.org/10.35292/ropj.v16i22.1018

Cerrillo i Martínez, A. (2021). La transparencia de los algoritmos que utilizan las administraciones públicas». *Anuario de Transparencia Local*, *3*, 41-78.

Corte Constitucional de Colombia. (2025). *Sentencia T-067 de 2025 (Exp. T-8.202.533)*. https://www.corteconstitucional.gov.co/relatoria/2025/T-067-25.htm

Gamero Casado, E. (2021). Enfoque europeo de inteligencia artificial. *Revista de Derecho Administrativo*, *20*, 268-289. https://revistas.pucp.edu.pe/index.php/derechoadministrativo/article/view/25212/23802

Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual (Indecopi). (2024, 12 de noviembre). *Resolución N.º 000151-2024-PRE/INDECOPI: Conformación del equipo de trabajo para la implementación de la inteligencia artificial en el INDECOPI*. https://www.gob.pe/institucion/indecopi/normas-legales/6178848-000151-2024-pre-indecopi

Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual (Indecopi). (2025a). *Resolución N.º 000062-2025-GEG/INDECOPI: Lineamientos para el uso ético de la inteligencia artificial*. https://www.gob.pe/institucion/indecopi/normas-legales/6822195-000062-2025-geg-indecopi

Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual (Indecopi). (2025b). *Resolución Final N.º 083-2025/CC3*. https://

img.lpderecho.pe/wp-content/uploads/2026/01/Expediente-083-2025CC3-LPDerecho.pdf

Piñar Mañas, J. (2008). ¿Existe privacidad? https://archivos.juridicas.unam.mx/www/bjv/libros/12/5669/3.pdf

Remolina Angarita, N. (2022, 21 de noviembre). Inteligencia artificial y tratamiento de datos personales. Ámbito Jurídico. https://www.ambitojuridico.com/noticias/analisis/educacion-y-cultura/inteligencia-artificial-y-tratamiento-de-datos-personales

Sánchez Acevedo, M. E. (2022). La inteligencia artificial en el sector público y su límite respecto de los derechos fundamentales. *Estudios Constitucionales*, *20*(2), 257-284. https://doi.org/10.4067/S0718-52002022000200257

Sánchez Vilanova, M. (2022). *El uso de algoritmos predictivos en el derecho penal: A propósito de la sentencia de la Corte de Justicia del Distrito de La Haya (Países Bajos) sobre SyRI, de 5 de febrero de 2020*. https://doi.org/10.36151/TD.2022.059

The Hague District Court. (2020, 5 de febrero). *SyRI (System Risk Indication) case*. https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-legislation-in-breach-of-European-Convention-on-Human-Rights.aspx

Tribunal de Justicia de la Unión Europea. (2023, 7 de diciembre). *Caso C-634/21, SCHUFA Holding AG*. https://curia.europa.eu/jcms/upload/docs/application/pdf/2023-12/cp230186es.pdf

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 173-189

INTELIGENCIA ARTIFICIAL EN LA IMPARTICIÓN DE JUSTICIA EN AMÉRICA LATINA

ARTIFICIAL INTELLIGENCE IN THE JUSTICE ADMINISTRATION IN LATIN AMERICA

Olivia Andrea Mendoza Enríquez

División de Estudios Jurídicos del Centro de
Investigación y Docencia Económicas (CIDE)

Resumen

Este artículo analiza el panorama actual del uso de la inteligencia artificial (IA) en los sistemas de justicia de América Latina. A través de un enfoque crítico, se examinan los beneficios de la automatización frente a los riesgos estructurales que la implementación de agentes inteligentes supone para los derechos humanos, entre los que se encuentran garantizar el debido proceso, la garantía del derecho a la no discriminación, la transparencia algorítmica y la brecha digital en materia de acceso a la justicia, entre otros.

El estudio identifica las tensiones éticas y legales derivadas del uso de IA en la toma de decisiones judiciales. Finalmente, se proponen los elementos jurídicos necesarios desde un enfoque preventivo para garantizar un despliegue seguro, ético y responsable de la IA.

El objetivo de este artículo es ofrecer una hoja de ruta normativa que armonice la eficiencia tecnológica con la protección de los derechos humanos en la región.

Palabras clave

Inteligencia artificial, acceso a la justicia, derechos humanos, sistemas de justicia, análisis de datos

Abstract

This article analyzes the context of artificial intelligence (AI) use in Latin American justice systems. Through a critical approach, it examines the benefits of automation versus the structural risks that the implementation of intelligent agents poses to human rights, including due process safeguard, the right to non-discrimination, algorithmic transparency, and the digital breach in access to justice, among others.

The study identifies the ethical and legal tensions arising from the use of AI in judicial decision-making. Finally, it proposes the necessary legal elements from a preventive perspective to ensure the safe, ethical, and responsible deployment of AI.

The aim of this document is to offer a regulatory roadmap that harmonizes technological efficiency with the protection of human rights in the region.

Keywords

Artificial intelligence, justice access, human rights, justice systems, data analysis

I. Introducción

El vertiginoso desarrollo tecnológico ha permeado en todos los ámbitos de la actividad humana. La aparición de Chat GPT hizo posible que el mundo se familiarizara con el concepto de la IA, aunque se tratara de un término que no era tan nuevo como parecía y que ya se usaba desde los años sesenta del siglo pasado[1].

La masificación del uso de la IA en el mundo ha permitido la incorporación de esta tecnología en casi todas las actividades cotidianas, eficientando el análisis de datos, la gestión de procesos y la toma de decisiones. No obstante, una preocupación permanente en el uso de agentes de IA, tiene que ver con los aspectos éticos y legales de su desarrollo, implementación y uso.

Para tener una mejor dimensión de los desafíos respecto del fenómeno de la IA en la actualidad, según lo reportado por el Índice Latinoamericano de Inteligencia Artificial, el fraude de identidad aumentó un 300 % y esto tiene una relación directa con la democratización y avance de la inteligencia artificial generativa y la tecnología *deepfake*[2].

Por otro lado, uno de los usos de IA que se está explorando en el mundo, es la incorporación de agentes inteligentes en los sistemas de justicia de los Estados nación. Su uso puede ir desde la gestión y la administración hasta la impartición de justicia, lo que requiere ciertos grados de autonomía y, por tanto, surgen algunas preocupaciones especialmente en el campo de los derechos humanos, entre las que se encuentran la necesidad de garantizar el derecho a la no discriminación, el acceso a la justicia pronta y expedita y el derecho a conocer cómo se toman las decisiones a través de algoritmos insertos en la IA[3].

El uso de inteligencia artificial aplicada a la impartición de justicia en América Latina se ha hecho por parte de Brasil (siendo el líder regional), Argentina, Colombia, México, Panamá y Chile[4].

Derivado de lo anterior, una de las preguntas eje de los últimos años se ha centrado en cómo hacer uso de la inteligencia artificial de manera segura y las posibles respuestas se discuten principalmente en foros económicos y regulatorios.

1 El término «inteligencia artificial» (IA) se utilizó formalmente por primera vez en 1956. Fue acuñado por el informático John McCarthy en la Conferencia de Dartmouth (oficialmente llamada *Dartmouth Summer Research Project on Artificial Intelligence*). Este evento es considerado el certificado de nacimiento de la disciplina como campo de estudio científico. Información disponible en: https://www.ibm.com/mx-es/think/topics/history-of-artificial-intelligence

2 Índice Latinoamericano de Inteligencia Artificial 2025. https://indicelatam.cl/wp-content/uploads/2025/10/Documento_ILIA_2025.pdf

3 Inteligencia artificial legal en Latinoamérica – ¿Qué países la usan en sus sistemas judiciales?. Disponible en: https://blog.lexlatam.ai/tecnologia-legal/inteligencia-artificial-legal-paises-latinoamerica/.

4 Inteligencia Artificial y su aplicación en los Sistemas de Justicia de América Latina, Instituto Belisario Domínguez del Senado de México, pp. 6-13. http://bibliodigitalibd.senado.gob.mx/bitstream/handle/123456789/5594/TE_101_IA_Sistemas_Justicia.pdf?sequence=1&isAllowed=.

En las siguientes líneas se analizarán algunos casos de uso de IA en el campo de la impartición de justicia, se nombrarán algunos riesgos asociados a la implementación de IA desde la perspectiva de los derechos humanos y se proponen los pilares que garantizan bajo un enfoque preventivo reducir los riesgos asociados al uso de la IA en la impartición de justicia, incorporando los beneficios que esta tecnología supone.

II. Casos de uso de la inteligencia artificial en la impartición de justicia en América Latina

Como se ha dicho en líneas previas, la masificación del uso de IA ha alcanzado espacios de la vida cotidiana y algunos de los sistemas judiciales de América Latina no han sido la excepción. La comparativa del uso de IA respecto a otras latitudes del mundo, pone en evidencia que el aprovechamiento regional de esta tecnología para la impartición de justicia se encuentra en etapa temprana[5], lo cual brinda a los actores del ecosistema digital la oportunidad de pensar cómo deberían ser e interactuar los agentes inteligentes, considerando un marco de los derechos humanos como obligación de los Estados nación.

Las bondades de tecnologías como la IA, permiten pensar en sistemas que solucionen problemas tradicionales asociados a la impartición de justicia, tales como la efectiva garantía de acceso a la justicia para todas las personas, resolver problemas como las altas cargas administrativas que tienen los servidores públicos con funciones judiciales, la falta de recursos y el rezago propio que la pandemia por covid-19 trajo a los sistemas de justicia en la región de América Latina.

Antes de hablar de casos concretos del uso de la IA en sistemas judiciales de América Latina, se analizarán algunos elementos a considerar respecto de la situación de evolución de la IA[6] y de las preocupaciones que surgen en el mundo, particularmente cuando estos agentes inteligentes se vuelven cada vez más autónomos y toman decisiones sobre derechos como la libertad de las personas y en general sobre los derechos humanos. Ejemplo de esto es la tendencia a ya no sistematizar procesos dentro de las organizaciones, sino sistematizar a través de agentes de IA, las decisiones que involucran derechos de las personas.

De esta situación surge un dilema entre la libertad y la pérdida de control de lo que se conoce como *human in the loop*, en el que el ser humano está involucrado en la toma de decisiones y la transición hacia agentes autónomos en los que la supervisión humana se vuelve secundaria o inexistente.

5 Preparación del sector judicial para la Inteligencia Artifical en América Latína, Centro de Estudios en Tecnología y Sociedad (CETyS) de la Universidad de San Andrés, Argentina, 2022, p. 6. Disponible en: https://cetys.lat/wp-content/uploads/2021/10/CompiladoV4.pdf

6 Lo que está pasando a nivel global en materia de IA está definiendo no solo los modelos de negocio por parte de las grandes tecnológicas, sino qué actores aprovecharán esos modelos, quién jugará el rol de países periféricos, quién aportará los datos para que tecnologías como la IA funcionen e incluso, qué actores absorberán el impacto ambiental que tiene la creciente demanda de IA en el mundo.

Lo que hemos visto en 2025 y 2026 anticipa una tendencia evolutiva: una relación estrecha entre la política y la economía global que incide en los modelos regulatorios y que define nuevas lógicas para el funcionamiento de la IA (indistintamente que se use para eficientar la logística de alguna corporación o que se use para tomar decisiones que inciden en la esfera de los derechos de las personas)[7].

En este sentido, en materia de IA existen dos modelos regulatorios predominantes en el mundo: el modelo regulatorio europeo y el anglosajón. El modelo regulatorio europeo o de derecho continental tiene un énfasis histórico en los derechos humanos, (reconociendo la dignidad humana como el límite del desarrollo tecnológico) y el segundo modelo, que es el de derecho anglosajón o *common law*, caracterizado por la gran influencia del mercado, la autorregulación y el establecimiento de límites en términos de la oferta y la demanda. Ambos han moldeado la arquitectura de la IA.

También es importante mencionar que existe una disparidad significativa entre el acceso a la IA y el aprovechamiento que se le da a esta tecnología: mientras existen países que desarrollan y explotan su propia tecnología, otros países son espectadores de procesos que los limitan y llevan a problemas como la falta de neutralidad tecnológica, el funcionamiento de tecnología no acorde a los marcos legales domésticos, la accesibilidad e imperativo tecnológico. Todo esto, supone un riesgo para los derechos humanos.

Dicho lo anterior, se puede advertir una relación directa entre la arquitectura de la IA, su desarrollo y evolución y la manera en la que acaba siendo utilizada en los sistemas de impartición de justicia. Mientras algunos países desarrollados hacen uso y aprovechamiento de la IA, otros países tendrán que esperar a la adquisición de estas herramientas tecnológicas, teniendo que superar desafíos que van desde la falta de datos para perfeccionar el entrenamiento de los agentes de IA, hasta limitaciones presupuestales del Estado, pasando también por la resistencia al paradigma tecnológico.

Dicho lo anterior, hablemos de algunos casos relevantes del uso de IA en sistemas de impartición de justicia en América Latina:

a) Prometea (Argentina): sin duda alguna se trata de uno de los casos más paradigmáticos del uso de IA para la impartición de justicia, al haber sido el primer sistema en ser desarrollado y utilizado en la región para tal fin. Este agente inteligente se desarrolló con el objetivo de agilizar procesos judiciales

7 Un ejemplo ilustrativo es lo sucedido en marzo de 2026 cuando el Pentágono declaró a la empresa Anthropic como un *supply chain risk* que se entiende como un riesgo para la cadena de suministro en la IA militar, por lo que se restringe su uso en contratos federales de Estados Unidos de Norteamérica, especialmente en proyectos relacionados con defensa. En la práctica, se trata de un veto tecnológico que podría afectar la presencia de la empresa en el ecosistema de seguridad nacional. Esta situación se presenta derivado de la negativa de dicha empresa a permitir el uso de su IA en vigilancia masiva y armas autónomas, particularmente en el marco de la guerra iniciada por Estados Unidos de Norteamérica contra Irán en marzo de 2026.

a través de la automatización de tareas rutinarias, reduciendo tiempos de gestión en más de un 90 % en casos concretos[8].

En palabras de su creador, el profesor Corvalán de la Universidad de Buenos Aires,

> el procedimiento es conducido íntegramente por la IA, de la siguiente manera: llega un expediente a dictaminar, que no ha sido analizado por ninguna persona. Se carga entonces el número de expediente a la inteligencia artificial Prometea, y en pocos segundos después pasa todo lo que se detalla a continuación. El sistema de IA busca la carátula en la página del Tribunal Superior de Justicia de la Ciudad de Buenos Aires, lo asocia con otro número (vinculado a las actuaciones principales) y luego va a la página del Poder Judicial de la Ciudad Autónoma de Buenos Aires (Juscaba). Busca y lee las sentencias de primera y segunda instancia, luego analiza más de 1400 dictámenes (emitidos durante 2016 y 2017), para finalmente emitir la predicción. En concreto nos dice que detecta un modelo determinado para resolver el expediente y nos ofrece la posibilidad de completar algunos datos para imprimir o enviar a revisar el dictamen con base en ese modelo (esto mismo podría hacer, si se tratara de dictar una sentencia)[9].

Prometea está asociado con un nuevo enfoque de trabajo, que procura aplicar IA en el desempeño de diferentes tareas vinculadas a la práctica del sistema de justicia. Se trata de un sistema de *software* que tiene como cometido principal la automatización de tareas reiterativas y la aplicación de IA para la elaboración automática de dictámenes jurídicos basándose en casos análogos para cuya solución ya existen precedentes judiciales reiterados. El diseño y la implementación del sistema se enmarcaron dentro del Plan Estratégico de la Fiscalía, como acciones proactivas para mejorar la eficiencia y la calidad del trabajo de la entidad[10].

b) PretorIA (Colombia): utilizada por la Corte Constitucional de Colombia para seleccionar, entre miles de tutelas (recursos de amparo), aquellas que cumplen con los requisitos para ser revisadas por los magistrados. PretorIA detecta patrones y clasifica las más de 3400 tutelas que recibe a diario la Corte Constitucional, con el propósito de contribuir a la identificación de los casos que se podrían seleccionar para su revisión. Además, PretorIA permite tener un panorama general del tipo de tutelas y los casos novedosos que llegan a la Corte[11].

PretorIA es un proyecto de inteligencia artificial en desarrollo, que tiene el propósito de hacer más eficiente el proceso de selección de los casos de tutela judicial de los derechos fundamentales.

8 Elsa Estevez, Sebastián T. Linares y Pablo Fillotrani, *Transformando la administración de justicia con herramientas de Inteligencia Artificial* (BID, 2020).

9 Juan Corvalán, *La primera inteligencia artificial predictiva al servicio de la Justicia: Prometea* (Thomson Reuters-La Ley, 2017).

10 Estevez, Linares y Fillotrani, *Transformando la administración de justicia…*

11 Sistema de Algoritmos Públicos, *PretorIA (anteriormente conocido como Prometea)*, s. f.

El caso de PretorIA es de interés por la entidad que lo desarrolla y por el contexto en que lo hace. Es un proyecto desarrollado por la Corte Constitucional para filtrar mejor los casos y optimizar su desempeño como máximo intérprete de la Constitución. Es de interés también como parte de los procesos de despliegue de tecnologías digitales en el sector justicia, tanto en su vertiente puramente técnica, como por los impactos en la organización[12].

c) JusticIA, Sor Juana y JuliIA (México): se trata de asistentes virtuales utilizados por la Suprema Corte de Justicia que permiten a la ciudadanía consultar criterios y datos oficiales mediante lenguaje natural.

En el caso de JusticIA, permite conocer información de temas judiciales publicada en la página de internet de la Suprema Corte de Justicia de la Nación en lenguaje natural para responder preguntas a través de respuestas predefinidas y en algunos casos a través de IA generativa, con el objetivo de facilitar el acceso a información a través de respuestas basadas en fuentes oficiales. Ayuda a localizar información sobre temas clave como: transparencia ciudadana y acceso a la información, justicia digital, obligaciones de transparencia y protección de datos personales, denuncias contra personas servidoras públicas del Alto Tribunal y criterios y resoluciones. Este asistente virtual fue diseñado bajo el principio de privacidad por diseño, lo que significa que no necesita recolectar datos personales para su funcionamiento. Además, la herramienta no utiliza cookies ni rastreo de navegación[13].

Para el caso de Sor Juana, se trata de una herramienta de inteligencia artificial desarrollada por la ponencia de la Ministra Ana Margarita Ríos Farjat, con apoyo de diferentes herramientas como Streamlit, Google, Pinecone, así como código propio, diseñada para facilitar la revisión, comprensión y socialización del contenido de las versiones públicas y sentencias[14]. De acuerdo con los datos estadísticos 6 mil personas ya han consultado contenido en la plataforma[15].

La IA tiene una función que abona a la justicia abierta y que permite que las personas conozcan el trabajo que realiza en Poder Judicial Federal y, con ello, conozcan y hagan valer sus derechos humanos.[16]

Por otro lado, JulIA es un proyecto piloto de un buscador jurídico también implementado por la Suprema Corte de Justicia de la Nación (SCJN) que incorpora

12 Víctor Saavedra y Juan Carlos Upegui, *PretorIA y la automatización del procesamiento de causas de derechos humanos* (Dejusticia, 2021), pp. 25-26.

13 Suprema Corte de Justicia de la Nación, *La Corte impulsa el asistente virtual "JusticIA" una herramienta que usa inteligencia artificial para la consulta de información relacionada con temas judiciales publicada en su portal web*, 27 de mayo de 2025.

14 Aplicación de Chatbot Sor Juana, desarollada por la ponencia de la Ministra Ríos Farjat: https://ponenciamamrfgpt.streamlit.app/

15 Presentan la Plataforma Sor Juana, un avance en materia de transparencia judicial, nota de prensa de la Suprema Corte de Justicia de la Nación: https://www.pluraltv.mx/noticias/detalle/presentan-plataforma-sor-juana-avance-materia-transparencia-judicial

16 Una conversación con Sor Juana en el siglo XXI: https://www.revistaabogacia.com/una-conversacion-con-sor-juana-en-el-siglo-xxi/

inteligencia artificial y fue presentado a finales de 2022[17]. Se trata de un sitio electrónico que involucra el uso de herramientas digitales, con el fin de transitar a un modelo de justicia digital. La plataforma permite buscar por el significado o el sentido de una oración en lenguaje coloquial, de manera que su uso es más sencillo que el buscador jurídico. La búsqueda emplea algoritmos de inteligencia artificial, lo que lleva una mayor precisión y a navegar en un universo de resultados. Esta plataforma es utilizada dentro de la Suprema Corte de Justicia de la Nación para identificar precedentes y su consecuencia jurídica.

Como se ha visto en líneas previas, se pueden identificar agentes de IA que sirven como buscadores (tanto de información que se ha suministrado a la plataforma previamente como de coincidencias abiertas a fuentes públicas), algunos otros que funcionan como *chatbots* de interacción y orientación para las personas interesadas y aquéllos que usan IA generativa en algún grado dentro del proceso.

De esto también se puede identificar que algunos agentes toman decisiones de forma autónoma y otros solo buscan coincidencias para mostrar resultados.

Para el caso de México por ejemplo, también es importante considerar que los avances en el uso de IA en materia de justicia, se dan principalmente a nivel federal y están sustentados por un interés de la Corte Suprema. No obstante, se advierte también que no existe una estrategia consensada de cómo incorporar la IA a nivel institucional ya que por ejemplo, algunos casos se tratan de esfuerzos de ministros en particular, sin lograr institucionalizar con una visión estratégica y transversal de una agenda para el uso seguro de IA. Esto es necesario, por ejemplo, para generar datos que perfeccionen el funcionamiento de los agentes inteligentes.

Otra cosa nada menor que se debe observar, es que algunos de los agentes de IA expuestos previamente, utilizan otros aplicativos de corporaciones como Google, lo cual deja sobre la mesa la pregunta sobre la armonización y concordancia de las políticas de privacidad de las empresas tecnológicas respecto de los marcos legales domésticos. Este punto también abre un debate respecto a la soberanía del dato y la soberanía tecnológica.

La disparidad respecto del alcance de los sistemas de IA para la impartición de justicia en América Latina es considerable, por ejemplo, para el caso de Prometea en Argentina, se presupuestó el ejercicio de recursos públicos para la creación del sistema. Para el caso de México, al no haber una estrategia acordada, la primera limitante tiene que ver con la falta de planeación presupuestal para el desarrollo a corto, mediano y largo plazo de agentes de IA. No obstante, en este rubro, se advierte como ha demostrado el caso de Prometea, que la inversión de dinero público tiene una tasa de retorno garantizada, al bajar las cargas administrativas de tareas necesarias para la impartición de justicia que se puedan sistematizar o automatizar.

También al analizarse los sistemas de IA más relevantes de la región, sería interesante pensar en una estrategia regional de datos abiertos, de justicia abierta, de interoperabilidad y del uso eficiente de tecnologías como la IA.

17 Sistema de Algoritmos Públicos. *JulIA*. s. f.

En conclusión, nos encontramos en una etapa temprana del uso de agentes de IA en la impartición de justicia a nivel regional, pese a los esfuerzos de los ejemplos citados previamente.

III. Riesgos de uso de los sistemas de inteligencia artificial desde una perspectiva de derechos humanos

Es conveniente iniciar este apartado hablando de las bondades del uso correcto de la tecnología. Considerar la posibilidad de frenar la incorporación tecnológica a las instituciones, simplemente representaría un desaprovechamiento de recursos humanos, financieros y materiales[18].

La pregunta no es si se usa o no la tecnología, la pregunta es cómo la incorporamos de manera eficiente y segura dentro de las organizaciones.

Para responder esta pregunta, es necesario identificar los principales riesgos asociados al desarrollo, implementación y uso de la IA desde la perspectiva de los derechos humanos, porque, aunque resulte evidente, se debe considerar que los usuarios finales de cualquier IA, somos las personas.

Algunos datos para entender el contexto del que partimos se encuentran en el Índice Latinoamericano de Inteligencia Artificial, el cual indica que el uso de la IA por parte de empresas privadas en México es bajo en comparación con otros países latinoamericanos, obteniendo un puntaje de 12.5 en el Índice, con un promedio regional de 25. Sin embargo, México se desempeña mejor en los avances del sector público, ocupando el tercer lugar en América Latina después de Argentina y Uruguay[19].

En la misma línea de los datos, para comprender el fenómeno respecto de los riesgos asociados al uso de IA, el fraude de identidad aumentó un 300 % por uso de IA, ya que existen paquetes de herramientas como *GenAI* que generan *deepfakes*, rostros y voces sintéticas en tiempo real. En este sentido, la consultora estadounidense Gartner destacó que más del 25 % de todas las interacciones con consumidores serán gestionadas de principio a fin por «agentes totalmente autónomos», que pueden ser blancos fáciles para la suplantación sin una infraestructura de identidad confiable[20].

18 Un modelo económico basado en análisis de tareas, aplicado al total de los aproximadamente 1,44 millones de trabajadores de la Administración pública española a escala local, autonómica y nacional, estima el siguiente potencial: para el 67 % de los trabajadores de la Administración pública (más de 960.000), la IA generativa podría mejorar entre el 10 % y la mitad de sus tareas, lo que hace factible su integración en los procesos diarios; para el 9 % de las ocupaciones (alrededor de 130.000 trabajadores), el potencial es aún mayor: podrían beneficiarse de la incorporación de la IA generativa en más de la mitad de sus tareas; para el 24 % restante (unos 345.000 trabajadores) el potencial es bajo, debido a la naturaleza irremplazable de sus tareas. Jorge Galindo, Manuel Hidalgo y Carlos Victoria, Álvaro Fernández y Javier Martínez Santos, *El impacto de la IA en el sector público español*, 2025.

19 Unesco, *México: evaluación del estadío de preparación de la inteligencia artificial*, 2024.

20 Las pérdidas que ha causado el mal uso de la IA en México: https://www.informador.mx/mexico/Las-perdidas-que-ha-causado-el-mal-uso-de-la-Inteligencia-Artifi-

El último informe *Top Fraud Trends* de TransUnion reveló que 1 de cada 6 consumidores estadounidenses perdió dinero a causa de fraudes digitales en el último año, mediante esquemas por correo electrónico, llamadas telefónicas o mensajes de texto, con una pérdida mediana de USD 2307[21].

La taxonomía de dominios de riesgos de la IA clasifica los riesgos así: discriminación, privacidad y seguridad, actores maliciosos, interacción persona-ordenador, desinformación, socioeconómico y ambiental[22].

El Consejo de la Unión Europea señala dentro de los riesgos la opacidad, la insuficiente rendición de cuentas, los riesgos para la seguridad, los sesgos y discriminación y la falta de vinculación con los derechos fundamentales[23].

Para la Organización para la Cooperación y el Desarrollo Económico, OCDE, los desafíos de implementación de IA son: falta de datos de calidad y capacidad para compartirlos, las brechas de habilidades, la falta de marcos de actuación y de directrices sobre el uso de la IA[24].

En términos del objetivo de este trabajo y el eje conductor que es el uso de la IA en sistemas de impartición de justicia en América Latina, se identifican los siguientes riesgos a considerar:

1. Tecnosolucionismo: Uno de los peligros es la idea de que la IA puede considerarse una panacea cuando en realidad es solo una herramienta. A medida que vemos más avances en IA, aumenta la tentación de aplicar la toma de decisiones basada en IA a todos los problemas sociales. Pero la tecnología a menudo crea problemas mayores al intentar resolver problemas menores. Por ejemplo, los sistemas que agilizan y automatizan la aplicación de los servicios sociales pueden volverse rápidamente rígidos y negar el acceso a migrantes u otras personas que quedan desamparadas.[25]

Entender que la tecnología es el medio y no el fin, resulta primordial para evaluar cuáles procesos pueden ser eficientados a través de agentes de IA y cuáles no. Habrá momentos procesales en los cuáles hasta ahora es imposible sustituir la intervención humana (ya sea de jueces o funcionarios judiciales), por ejemplo en la evaluación probatoria, en la interpretación (sentencias con perspectiva de género acordes a un caso concreto) o en la misma sentencia, que

cial-20250819-0135.html

21 El aumento de fraudes con IA multiplica riesgos para usuarios y empresas estadounidenses: https://www.infobae.com/estados-unidos/2026/04/19/el-aumento-de-fraudes-con-ia-multiplica-riesgos-para-usuarios-y-empresas-estadounidenses/

22 Peter Slattery, Alexander Saeri, Emily Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper y Neil Thompson, *The Al Risk Repository A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence* (MIT, 2026).

23 Ventajas y riesgos de la IA, Consejo de la Unión Europea: https://www.consilium.europa.eu/es/policies/benefits-and-risks-of-ai/#risks

24 OCDE, *Gobernar con la Inteligencia Artificial. Panorama actual y hoja de ruta en las funciones centrales de gobierno*, 2025.

25 Stanford University. *SQ10. What are the most pressing dangers of AI?* 2021.

si bien puede ser elaborada y propuesta por un agente inteligente, requiere de la indispensable revisión y determinación humana.

2. Sesgo algorítmico y discriminación: Las personas no pueden ser solo una estadística judicial. Un riesgo asociado al uso de agentes de IA en la impartición de justicia está relacionado con las decisiones automatizadas que funcionan a través de algoritmos, que de la experiencia se ha visto, replican y maximizan prejuicios existentes en el mundo físico. Por ejemplo, recordemos el caso COMPAS[26] en Estados Unidos de Norteamérica, en el que el sistema calificaba de manera errónea a acusados afroamericanos otorgándoles una calificación de alto riesgo de reincidencia delictiva respecto de las personas blancas.

Las decisiones tomadas por IA no pueden ser consideradas un punto final y bajo un enfoque preventivo, existen figuras jurídicas que atenuarían riesgos asociados a la vulneración de derechos humanos derivada del uso de agentes inteligentes como el derecho de revisión frente a decisiones automatizadas y el derecho a no ser sometido a las mismas.

3. Falta de transparencia algorítmica: Este apartado tiene una relación directa con la discriminación algorítmica. En este sentido, a nivel normativo existen pocas disposiciones que obliguen a los actores que usan IA, a garantizar la transparencia algorítmica. Cuando en la región de América Latina se habla de dicha transparencia, se traspolan argumentos como la innovación y la propiedad intelectual que hay detrás del desarrollo algorítmico (secretos comerciales de los desarrolladores de *software*) y la idea de que la regulación frena el desarrollo científico.

No obstante lo anterior, la transparencia algorítmica no solo es elemento habilitador de la IA, sino una garantía de confianza en donde las personas sepan exactamente cómo se toman las decisiones y tengan la posibilidad de acceder a mecanismos legales de aclaración y rectificación en caso de que exista alguna imprecisión.

La falta de transparencia en los sistemas inteligentes aplicados a la justicia, pone en riesgo el derecho de defensa de los acusados y la garantía del debido proceso.

Desde la experiencia europea, tenemos algunos ejemplos importantes de cómo puede garantizarse la transparencia algorítmica, cuidando la innovación y los esfuerzos financieros que existen detrás de la investigación que permite nuevos productos comerciales. En este sentido, en 2023 se puso en marcha el Centro Europeo para la Transparencia Algorítmica (ECAT) que sirve para proporcionar conocimientos científicos y técnicos que apoyen la aplicación de la Ley

26 El Caso Wisconsin vs. Loomis es un hito legal donde el tribunal determinó que el uso de COMPAS no violaba el debido proceso, siempre que no fuera el único factor en la sentencia, aunque se criticó la falta de transparencia del algoritmo por ser privado. Véase ¿Cómo en Estados Unidos las matemáticas te pueden meter en prisión? Véase ¿Cómo en Estados Unidos las matemáticas te pueden meter en prisión? https://www.bbc.com/mundo/noticias-37679463.

de Servicios Digitales (DSA)[27] y la investigación sobre el impacto de los sistemas algorítmicos utilizados por las plataformas en línea y los motores de búsqueda. También esta institución investiga el impacto a largo plazo de los sistemas algorítmicos para fundamentar la formulación de políticas y contribuir al debate público[28].

En conclusión, el tema de transparencia algorítmica debe ser visto como parte necesaria de un enfoque preventivo que ayude a los actores del ecosistema digital a minimizar los riesgos asociados a la IA y por tanto, que las personas tengan más confianza y gocen de los beneficios del desarrollo tecnológico.

4. Seguridad de la información y políticas de ciberseguridad: las instituciones públicas suelen no tener una hoja crítica en materia de seguridad de la información en la IA.

También se advierte que dicha ruta, esté armonizada con una política correcta en materia de ciberseguridad. Para el caso de los agentes inteligentes para la impartición de justicia, se advierte que las estrategias organizacionales, las hojas críticas y los planes de transición e implementación tecnológica, deben estar sustentados primero en un marco constitucional y de derechos humanos, segundo, que deben considerar normativas complementarias como aquéllas en materia de archivos, transparencia judicial, justicia abierta, generación de datos abiertos y todo lo anterior, buscando el objetivo de plataformas interoperables que hagan más sencilla la función judicial tanto para servidores públicos como para los justiciables.

5. Brecha de datos, privacidad y protección de datos personales: uno de los riesgos más evidentes el uso de IA en la impartición de justicia es el tema del manejo de la información, la garantía del derecho a la privacidad y el cumplimiento de principios y deberes en materia de protección de datos personales. Para la mayor parte de países en América Latina, ya existen marcos jurídicos locales que orientan los esfuerzos y las obligaciones legales tanto del derecho a la privacidad como el de protección de datos personales. No obstante lo anterior, la situación se vuelve compleja cuando el servicio de IA que se contrata es prestado por una empresa trasnacional que no reconoce el marco jurídico nacional en materia de protección de datos personales como obligatorio, ni la competencia y jurisdicción de las autoridades en la materia. En otras palabras, a pesar de tener marcos legales robustos en materia de protección de datos personales, estos no siempre inciden en el diseño de los términos y condiciones de los servicios prestados por empresas como Open AI, Google o Meta.

6. Crisis del derecho internacional: Como hemos visto durante los últimos meses, el derecho internacional enfrenta una crisis de prevalencia, confianza y

27 La Ley de Servicios Digitales de 19 de octubre de 2022, es el marco regulatorio integral de la Unión Europea para intermediarios y plataformas en línea. Su objetivo es crear un entorno digital más seguro, proteger los derechos fundamentales de los usuarios y frenar la difusión de contenidos ilícitos, productos falsificados y desinformación. https://digital-strategy.ec.europa.eu/es/policies/digital-services-act

28 Acerca de ECAT, el Centro Europeo para la Transparencia Algorítmica. Información disponible en: https://algorithmic-transparency.ec.europa.eu/about_en.

fortalecimiento de las instituciones y de reglas como las del derecho internacional humanitario (que también aplican en algunos puntos a agentes inteligentes como las armas autónomas de destrucción masiva).

Si bien después de la Segunda Guerra Mundial los Estados se comprometieron a construir reglas mínimas que garantizaran los mismos derechos para todas las personas en un reconocimiento inédito de la dignidad humana, lo que hemos es que esos consensos, reglas e instituciones no son suficientes para por ejemplo, garantizar que la IA no se utilice en contra de las personas como es el caso de los drones no tripulados que toman decisiones autónomas y que en 2020 dispararon en contra de la población en Libia[29].

Todo lo anterior, termina suponiendo un riesgo para el uso seguro de la IA en la impartición de justicia, porque no hay compromisos internacionales mínimos (tanto corporativos como de normas) que eviten el riesgo de desarrollo de IA que afecte los derechos de las personas[30].

IV. Algunas propuestas para la implementación y uso seguro de la inteligencia artificial en la impartición de justicia

1. Voluntad institucional y creación de estrategias y hojas de ruta que cuenten con el consenso de todas las áreas involucradas. Identificar beneficios del uso de IA y traducirlo al impacto que esto tendría a las tareas sustanciales judiciales.

2. Estrategias y rutas críticas: A nivel institucional, contar con una estrategia de digitalización y generación de datos con valor (política de datos abiertos). Establecer las condiciones de desarrollo institucional, normativo y tecnológico.

3. Análisis de tareas y función judicial para determinar qué partes procesales son susceptibles de ser hechas por agentes de IA y cuáles deben ser hechas por humanos.

4. Enfoque preventivo y no reactivo. Se deberán instrumentar todas las medidas preventivas que eviten afectar los derechos de las personas. Optar por agentes inteligentes que estén hechos bajo enfoques de privacidad por diseño, transparencia algorítmica, certificaciones y auditorías de funcionamiento y en general, buena reputación corporativa (por ejemplo que tenga protocolos adecuados en casos de *data breach*).

5. Limitación clara contractual del régimen de responsabilidad entre los actores del ecosistema digital. Cláusulas contractuales que identifiquen los grados de responsabilidad en todas las etapas de desarrollo, implementación y uso

29 La ONU informa del primer ataque de drones autónomos a personas. Información disponible en: https://elpais.com/tecnologia/2021-05-28/la-onu-informa-del-primer-ataque-de-drones-autonomos-a-personas.html.

30 Amnistía Internacional Global: La vergonzosa decisión de Google de revocar su prohibición de usar inteligencia artificial para armas y vigilancia, un duro golpe para los derechos humanos: https://www.amnesty.org/es/latest/news/2025/02/global-googles-shameful-decision-to-reverse-its-ban-on-ai-for-weapons-and-surveillance-is-a-blow-for-human-rights/

de agentes inteligentes. Cláusulas como la portabilidad y la accesibilidad de la información que sea suministrada por una institución a la empresa que comercialice o preste el servicio de IA, resultan primordiales para dar certeza jurídica sobre un elemento tan sensible y pieza clave dentro de los poderes judiciales, como es la información judicial.

6. Análisis del uso de agentes de IA y la naturaleza de la información para determinar si las fuentes de información que usen los sistemas serán abiertas o cerradas. Esto debe considerar el riesgo asociado a las alucinaciones que se dan en la IA cuando se trata de entornos no controlados y/o abiertos y la responsabilidad que tienen las organizaciones respecto de información que dichos agentes den a través de canales institucionales[31].

7. Cuidar las contrataciones y subcontrataciones de *software* respecto de los datos propietarios que evalúan la conducta y deciden el futuro de las personas. En otras palabras, se deben habilitar mecanismos preventivos de riesgos en la estimación algorítmica de riesgo de un individuo en la sociedad.

8. Cuidar como esencia de los agentes inteligentes que en todo momento funcionen garantizando el debido proceso y señalar mecanismos de prevención de sesgos y vías de corrección.

9. Garantizar siempre la transparencia algorítmica que permita también cumplir con principios asociados a la IA como la explicabilidad y la trazabilidad de las decisiones tomadas por agentes inteligentes.

10. Cuidado en la calidad y limpieza del dato para no utilizar variables que permitan deducciones que deriven en vulneraciones al derecho de no discriminación.

11. Establecer mecanismos no solo de auditoría algorítmica sino de monitoreo permanente, especialmente frente al tratamiento de datos sensibles como la raza, el género y la condición social de las personas.

12. Supervisión humana o *human-in-the-loop* para que las decisiones finales que afecten derechos humanos de las personas sean tomadas por personas y no por agentes de IA.

13. Reconocimiento normativo de figuras legales que permitan que las personas tengan el control de decisiones tomadas por IA que les puedan afectar, por ejemplo, el recurso de revisión frente a decisiones automatizadas.

31 Por ejemplo, el caso de Air Canada y su chatbot que proporcionó información incorrecta a un viajero. La compañía utilizó el argumento de no responsabilidad sobre información dada por un agente inteligente, declarando que el chatbot era una "entidad jurídica independiente responsable de sus propios actos".

El Tribunal de Resolución Civil de Columbia Británica rechazó ese argumento y dictaminó que Air Canada debía pagar al pasajero que inició la reclamación, 812,02 dólares (642,64 libras esterlinas) en concepto de daños y perjuicios y costas judiciales. Véase Airline held liable for its chatbot giving passenger bad advice-what this means for travellers: https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know.

Finalmente, cuando hablamos de asimetrías en el desarrollo de la IA suelen venir acompañadas por la necesidad de protección y efectivización de derechos también dispares. Es decir, se deben afrontar desafíos vinculados a cuestiones de primera necesidad (acceso al agua, acceso a servicios esenciales, etc.), pero también es importante avanzar en proteger otras violaciones masivas de derechos en el mundo digital[32].

V. Conclusiones

Como hemos visto, el término de IA no es nuevo, a pesar de que se hizo popular con la aparición del Chat GPT. En este sentido, la preocupación actual se centra en la democratización del uso de la IA generativa, lo cual supone en algunos casos riesgos para los derechos de las personas, ya que las soluciones de seguridad informática tradicionales (por ejemplo, las medidas de autenticación financiera a través de la voz), han sido rebasadas por las posibilidades que brinda este tipo de tecnología.

Elementos geopolíticos que se viven en el panorama internacional, están influyendo en las arquitecturas tecnológicas y comerciales de los sistemas de IA. En este punto, el enfoque regulatorio por el que opten los Estados nación respecto de la IA, será fundamental para determinar la prevalencia entre mercado y los derechos de las personas, considerando que el uso masivo de la IA, diluyen cada vez más la línea entre política, economía y arquitectura de la tecnología.

La IA se ha incorporado a muchos procesos cotidianos de las personas y el ámbito judicial no ha sido la excepción con los casos analizados de México, Colombia y Argentina. Se advierte la necesidad de no perder el control humano sobre las decisiones, pese a que puede ser una tecnología muy efectiva para eficientar los procesos.

Se identifica que no puede pausarse el uso de la IA, sino pugnar por mecanismos preventivos que permitan un uso seguro de esta tecnología. En este rubro, el enfoque regulatorio será determinante para establecer salvaguardas efectivas de protección a las personas que permitan un uso seguro de la tecnología.

Los riesgos asociados a agentes inteligentes en materia de impartición de justicia están focalizados y pueden ser prevenidos desde un espacio jurídico y tecnológico.

Existen algunas directrices recomendables para el despliegue seguro de IA en el ámbito judicial, entre las que se encuentran: analizar procesos judiciales susceptibles de ser realizados por IA, asegurar la calidad y mejora de los datos, tener rutas estratégicas para el uso de IA a nivel institucional, garantizar el cumplimiento normativo secundario que involucra el funcionamiento de IA como las leyes en materia de protección de datos personales y archivos, contar con una política de ciberseguridad adecuada, etc.

32 Juan Corvalán, «Inteligencia artificial: retos, desafíos y oportunidades. Prometea: la primera inteligencia artificial de Latinoamérica al servicio de la Justicia», *Revista de Investigações Constitucionais*, vol. 5, n.º 1 (2018).

En general, lo que hemos visto en las líneas anteriores es una tecnología cuyo uso se ha masificado y nos lleva a pensar en formas equilibradas de aprovechamiento de uso, pero al mismo tiempo no logra resolverse la salvaguarda de los derechos humanos como la privacidad, la protección de datos personales y el derecho a la no discriminación.

Referencias bibliográficas

«Acerca de ECAT, el Centro Europeo para la Transparencia Algorítmica». En *Centro Europeo para la Transparencia Algorítmica*. https://algorithmic-transparency.ec.europa.eu/about_en

Aguilar Hernández, Claudia I. «Una conversación con Sor Juana en el siglo XXI». En *Abogacía*, 19 de noviembre de 2024. https://www.revistaabogacia.com/una-conversacion-con-sor-juana-en-el-siglo-xxi/

AO. «Las pérdidas que ha causado el mal uso de la IA en México». En *Informador*, 23 de agosto de 2025. https://www.informador.mx/mexico/Las-perdidas-que-ha-causado-el-mal-uso-de-la-Inteligencia-Artificial-20250819-0135.html.

«Aplicación de Chatbot Sor Juana, desarrollada por la Ponencia de la Magistrada Ríos Farjat». En *Ponencia Ríos Farjat*. https://ponenciamamrfgpt.streamlit.app/

«Global: La vergonzosa decisión de Google de revocar su prohibición de usar inteligencia artificial para armas y vigilancia, un duro golpe para los derechos humanos». En *Amnistía Internacional*, 6 de febrero de 2025. en: https://www.amnesty.org/es/latest/news/2025/02/global-googles-shameful-decision-to-reverse-its-ban-on-ai-for-weapons-and-surveillance-is-a-blow-for-human-rights/

«La historia de la IA». En *IBM think Topics*. https://www.ibm.com/mx-es/think/topics/history-of-artificial-intelligence.

Maybin, Simon. «¿Cómo en Estados Unidos las matemáticas te pueden meter en prisión?». En *BBC*, 17 de octubre de 2016. https://www.bbc.com/mundo/noticias-37679463

Murillo, Laura. «Presentan la Plataforma Sor Juana, un avance en materia de transparencia judicial, nota de prensa de la Suprema Corte de Justicia de la Nación». En *Plural TV*, 6 de diciembre de 2024. https://www.pluraltv.mx/noticias/detalle/presentan-plataforma-sor-juana-avance-materia-transparencia-judicial.

Ortiz, Osvaldo, «El aumento de fraudes con IA multiplica riesgos para usuarios y empresas estadounidenses». En *INFOBAE*, 19 de abril de 2026. https://www.infobae.com/estados-unidos/2026/04/19/el-aumento-de-fraudes-con-ia-multiplica-riesgos-para-usuarios-y-empresas-estadounidenses/.

Rodríguez, J.M. y Barrón, M.A. «Inteligencia Artificial y su aplicación en los Sistemas de Justicia de América Latina». En *Instituto Belisario Domínguez del Senado de la República*, marzo de 2022. http://bibliodigitalibd.senado.gob.mx/bitstream/handle/123456789/5594/TE_101_IA_Sistemas_Justicia.pdf?sequence=1&isAllowed=y

Roxana. «Inteligencia artificial legal en Latinoamérica – ¿Qué países la usan en sus sistemas judiciales?». En *Blog Lexlatam*, 16 de septiembre de 2025. https://blog.lexlatam.ai/tecnologia-legal/inteligencia-artificial-legal-paises-latinoamerica/.

Vega, Guillermo. «La ONU informa del primer ataque de drones autónomos a personas». En *El País*, 28 de mayo de 2021. https://elpais.com/tecnologia/2021-05-28/la-onu-informa-del-primer-ataque-de-drones-autonomos-a-personas.html

«Ventajas y riesgos de la IA, Consejo de Europa». En *Consejo de la Unión Europea*. https://www.consilium.europa.eu/es/policies/benefits-and-risks-of-ai/#risks.

Yagoda, María. «Airline held liable for its chatbot giving passenger bad advice - what this means for travellers». En *BBC*, 23 de febrero de 2024. https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 190-205

IA PARA LA VIGILANCIA AMBIENTAL: UN MODELO DE *CROWDSOURCING* E IA CON INCENTIVOS PARA LA PROTECCIÓN DE ÁREAS NATURALES PROTEGIDAS

AI FOR ENVIRONMENTAL SURVEILLANCE: A CROWDSOURCING AND AI MODEL WITH INCENTIVE MECHANISMS FOR THE PROTECTION OF NATURAL PROTECTED AREAS

Luis Enrique Vázquez de la Paz, Regina Coeli Segura Cardona, Andrea Anabell Barba González, María Fernanda Muñoz Ayala

Doctorandos en Derecho, Instituto Tecnológico
y de Estudios Superiores de Monterrey

Resumen

La vigilancia de las Áreas Naturales Protegidas (ANP) se enfrenta a la limitación física de los cuerpos de inspección frente a vastas extensiones territoriales. Este artículo propone un modelo disruptivo de fiscalización administrativa asistida que integra a la ciudadanía (visitantes habituales) como co-generadores de información para el proceso sancionador. El sistema se basa en una arquitectura de denuncia pseudoanónima que utiliza inteligencia artificial para el filtrado y validación de evidencia digital. Como elemento central, se introduce un mecanismo de incentivos vinculados a resultados directos de conservación, transformando la denuncia en una acción de participación proactiva. Mediante una metodología de *design science research*, se analiza la viabilidad técnica de automatizar la cadena de procesamiento de la información ciudadana hasta la generación de la sanción administrativa. El estudio concluye que este modelo no solo optimiza la capacidad operativa del organismo público, sino que democratiza la justicia ambiental y reduce la impunidad en zonas de difícil acceso.

Palabras clave

Participación ciudadana, *crowdsourcing*, inteligencia artificial, áreas naturales protegidas, sanciones administrativas

Abstract

Environmental Surveillance of Natural Protected Areas (NPAs) faces the physical constraints of the inspection units when confronted with vast territorial mases. This article proposes a disruptive model of Assisted Administrative Oversight that integrates the citizenry (regular visitors) as information co-generators for the sanctioning process. The system is predicated on a pseudo-anonymous reporting architecture that utilizes Artificial Intelligence for the filtering and validation of digital evidence. Central to this model is the introduction of an incentive mechanism linked to direct conservation outcomes, thereby transforming the act of reporting into proactive participation. Through a Design Science Research methodology, the technical feasibility of automating the processing chain—from citizen-sourced data to the generation of administrative sanctions—is analyzed. The study concludes that this model not only optimizes the operational capacity of public agencies but also democratizes environmental justice and mitigates impunity in remote or inaccessible regions.

Keywords

Civic participation, *crowdsourcing*, artificial intelligence, protected natural areas, administrative sanctions

I. Introducción

México constituye un referente en diversas áreas y temáticas. Desafortunadamente, la gestión y protección de las Áreas Naturales Protegidas (ANP) no se cuenta entre ellas. Por el contrario, el país atraviesa una profunda crisis de gestión ambiental: pese a los discursos y narrativas oficiales, la realidad se aparta sensiblemente de aquéllos y los daños resultan cada vez más cuantiosos.

Los organismos públicos en México, en particular los organismos públicos descentralizados encargados de la gestión ambiental, despliegan, sin lugar a duda, su mejor esfuerzo con los limitados recursos con los que cuentan. La precariedad presupuestal constituye una limitación estructural ineludible que genera daños materiales al patrimonio natural y, a su vez, alimenta un círculo vicioso de apatía ciudadana frente a la autoridad ambiental (Castro Salazar *et al*., 2021).

Autores como Machlis y Tichnell (1985) han llegado, incluso, a acuñar el concepto de *parques de papel*: espacios geográficos delimitados y legalmente constituidos, pero materialmente vulnerables ante actividades ilícitas como la tala clandestina, el cambio de uso de suelo y, sobre todo, los incendios provocados. La brecha entre los polígonos por vigilar y los recursos humanos disponibles para hacerlo constituye uno de los principales factores estructurales de la impunidad ambiental que impera en el país.

Jalisco, con el Bosque La Primavera (principal pulmón del área metropolitana de Guadalajara), constituye la ejemplificación más nítida de lo anterior, con una extensión superior a las 30,000 hectáreas (Organismo Público Descentralizado Bosque La Primavera, 2023). El OPD Bosque La Primavera enfrenta una presión urbana descomunal. De acuerdo con el programa de manejo del fuego, su estructura orgánica revela que las funciones de vigilancia y protección recaen en un personal operativo notoriamente reducido frente a las múltiples amenazas que confluyen sobre dicho territorio (Organismo Público Descentralizado Bosque La Primavera, 2024).

Esta limitación material tiene consecuencias devastadoras. Diversos reportes académicos de la Universidad de Guadalajara documentan que los incidentes más recientes han afectado hasta el 8% del área natural protegida por evento catastrófico (González Cuevas, citado en El Informador, 2023; OPD Bosque La Primavera, 2023; Valdez-Rosas, 2023). La falta de recursos y de apoyo institucional en la materia ha generado no sólo limitaciones en la mitigación de los incendios, sino también, de manera particularmente preocupante, serias dificultades en la identificación de los responsables (Jardel, citado en Mongabay, 2023; Valdez-Rosas, 2023).

Los reportes semanales de incendios forestales de la Comisión Nacional Forestal confirman una tendencia clara: la mayoría de los incidentes son detectados cuando ya han alcanzado proporciones críticas, lo que evidencia que la inspección tradicional (basada en recorridos físicos programados) resulta insuficiente frente a la dinámica del fuego y al accionar de grupos infractores (Conafor, 2024; Carpio-Domínguez, 2024). La evidencia empírica disponible refuerza esta tesis. Diversos estudios recientes señalan cómo el organismo encargado de la fiscalización, al combinar una estructura centralizada con una grave escasez de

inspectores por kilómetro cuadrado, genera un cuello de botella administrativo de proporciones críticas (Castro Salazar *et al.*, 2021; Carpio-Domínguez, 2024; Profepa, s. f.).

Según los indicadores de fiscalización, el tiempo que transcurre entre el ingreso de una denuncia ciudadana convencional y la presencia de un inspector en el sitio puede ser de días o incluso semanas; plazo suficiente para que la evidencia que podría vincular a los responsables del daño ambiental se pierda por completo (Castro Salazar *et al.*, 2021; Cervantes, Martínez y Núñez, 2023). Bajo este panorama, la justicia ambiental en las áreas naturales protegidas es, por tanto, una justicia reactiva y no preventiva, cuya eficacia queda acotada por la capacidad operativa de unos cuerpos de seguridad insuficientes frente a la inmensidad geográfica del territorio (Castro Salazar *et al.*, 2021; Carpio-Domínguez, 2024).

Adicionalmente, la falta de incentivos para la vigilancia comunitaria formalizada ha dejado a los visitantes habituales (senderistas, ciclistas y residentes locales) como meros observadores pasivos del deterioro que sufre el Bosque La Primavera. Las entrevistas y notas periodísticas que dan cobertura a los diversos incendios sugieren que, aunque la ciudadanía reporta a través de redes sociales y llamadas de emergencia, dichos datos difícilmente se convierten en evidencia procesal aprovechable para un procedimiento administrativo sancionador o para la integración de una carpeta de investigación penal.

Ante la evidente imposibilidad de generar un aumento presupuestal de la magnitud necesaria para cubrir la complejidad y amplitud del territorio de las áreas naturales protegidas, resulta indispensable recurrir a medidas innovadoras que combinen el uso de tecnología con el apoyo de las comunidades para mejorar la gestión ambiental. Mientras el modelo de vigilancia continúe dependiendo de la presencia física de un funcionario público para validar cada infracción en extensiones de miles de hectáreas, el resultado seguirá siendo la degradación progresiva del patrimonio natural. El modelo ha alcanzado, evidentemente, sus límites operativos, lo que vuelve imperativa una transición hacia alternativas como un sistema de fiscalización distribuido y tecnológicamente asistido.

II. Tecnología en la gestión ambiental de la ANP

La incorporación de tecnología en la administración pública ha demostrado, en todas sus interacciones, que los apoyos digitales mejoran sustancialmente la prestación de servicios públicos. Tanto a nivel local como nacional e internacional, son numerosos los ejemplos en que la tecnología ha irrumpido en el sector público generando resultados notorios en beneficio de la población (OCDE & CAF, 2023; Criado, 2021).

En efecto, la evolución de la administración pública pasa necesariamente por su interrelación y sinergia con la tecnología. La digitalización forma parte de la reingeniería estructural que todo gobierno requiere. El caso del Servicio de Administración Tributaria (SAT) de la Secretaría de Hacienda y Crédito Público de México ilustra con claridad que la cuestión no reside en el tamaño o

la complejidad operacional del organismo, sino en la determinación institucional y en la calidad de la implementación (CETYS, 2019; LexLatin, 2026).

Basta evidenciar cómo el SAT ha logrado un crecimiento sostenido de su recaudación gracias a la tecnología: a partir de las reformas introducidas hace más de una década en materia de contabilidad electrónica (SAT, 2014) y de la habilitación de herramientas y complementos para facilitar dicha obligación al contribuyente, las compulsas fiscales y las auditorías masivas gestionadas por un solo operador son hoy una realidad consolidada que representa un éxito contundente del proceso de digitalización de la administración tributaria (CETYS, 2019; Facturando.mx, 2026).

En un contexto más reciente, destaca el caso de la Agencia Nacional de Energía Eléctrica de Brasil, que desarrolló el Geospatial Transmission Management System (GTMS), herramienta que emplea inteligencia artificial para el análisis de imágenes satelitales con el propósito de prevenir apagones causados por incendios forestales próximos a las líneas de transmisión eléctrica. Gracias a este sistema, la autoridad reguladora puede emitir alertas preventivas y orientar inspecciones de manera más precisa, reduciendo la necesidad de supervisiones físicas en campo. Como resultado, el GTMS ha permitido disminuir aproximadamente en un 89 % los apagones asociados a incendios forestales en las líneas monitoreadas, fortaleciendo la confiabilidad del sistema eléctrico nacional (OECD Observatory of Public Sector Innovation, 2024a).

En Ámsterdam, la inteligencia artificial se utiliza para mejorar la accesibilidad urbana y facilitar el acceso a servicios sociales mediante herramientas digitales que analizan rutas accesibles para personas con movilidad reducida y optimizan la prestación de servicios públicos. Estas iniciativas forman parte de la estrategia municipal de innovación responsable, orientada a mejorar la sostenibilidad ambiental, la movilidad urbana y la inclusión social (City of Amsterdam, 2024).

Asimismo, el proyecto vCity, impulsado por el Barcelona *Supercomputing Center* en colaboración con gobiernos locales como el de Barcelona, desarrolla plataformas de gemelos digitales urbanos que permiten simular el impacto de decisiones públicas antes de implementarlas. Estas herramientas contribuyen a abordar desafíos urbanos complejos como el cambio climático, la desigualdad social y la movilidad sostenible, incorporando además mecanismos de participación ciudadana a través de plataformas digitales como Decidim (OECD Observatory of Public Sector Innovation, 2024e).

Finalmente, el caso más significativo de integración tecnológica a nivel estatal es el de Tallinn, Estonia, cuya capital se ha consolidado como referente mundial de la digitalización gubernamental: prácticamente la totalidad de sus trámites se encuentran disponibles a través de su plataforma de *e-Government* bajo el principio *once-only*, conforme al cual la administración no puede requerir al ciudadano información que ya obra en sus propios registros (e-Estonia, 2024). Este ecosistema digital, en el que la identidad ciudadana está protegida mediante tecnología de cadena de bloques (*blockchain*) y la burocracia presencial

ha sido prácticamente erradicada, ha generado ahorros cuantificables tanto en horas de productividad como en costos operativos para el Estado.

La experiencia de Tallin, sumada a los casos de México y Brasil, confirma que la incorporación de la tecnología en la gestión pública no es una opción de modernización, sino una necesidad imperativa. En última instancia, la narrativa del progreso gubernamental contemporáneo se escribe con código y datos, consolidando un Estado que es, por primera vez, tan dinámico como la sociedad a la que sirve.

1. El cambio de paradigma: de la vigilancia vertical a la red social de conservación

La crisis de supervisión del Organismo Público Descentralizado (OPD) Bosque La Primavera se fundamenta en una desproporción crítica entre la extensión del territorio bajo su responsabilidad y el capital humano disponible para gestionarlo. Con una superficie de 30,500 hectáreas caracterizada por una alta complejidad operativa (derivada, entre otros factores, de sus 22 puntos de acceso), la Dirección de Protección y Vigilancia cuenta con apenas 28 elementos, lo que representa una carga de 1,089 hectáreas por persona. Esta precariedad se agudiza al analizar específicamente el cuerpo de guardabosques: solo 10 efectivos son responsables de cubrir 3,050 hectáreas cada uno, asumiendo simultáneamente tareas de detección de ilícitos y de primera respuesta ante incendios, lo que genera inevitables zonas ciegas operativas de difícil subsanación bajo el modelo de vigilancia vigente (Huerta-Martínez e Ibarra-Montoya, 2014).

A la carencia de personal se suma un grave déficit en la infraestructura logística que imposibilita una vigilancia territorial efectiva. Actualmente, la capacidad de desplazamiento se encuentra severamente limitada: de las 9 unidades vehiculares asignadas al OPD, únicamente 5 están operativas, lo que obliga a cada vehículo funcional a cubrir una extensión aproximada de 6,100 hectáreas (Organismo Público Descentralizado Bosque La Primavera, 2024). Esta debilidad estructural (administrativa y técnica a la vez) condiciona de manera directa los tiempos de respuesta y la frecuencia de los recorridos de vigilancia, corroborando que el modelo de fiscalización tradicional resulta manifiestamente insuficiente para garantizar la integridad territorial de una de las áreas naturales protegidas más relevantes del estado de Jalisco.

Este problema requiere, en primer término, del apoyo de la tecnología para optimizar recursos y multiplicar la capacidad operativa; no obstante, dada la magnitud del reto, resulta necesario un escalón adicional. La vía más razonable para lograr una cobertura efectiva en el corto plazo pasa por la implementación de un modelo de gobernanza y fiscalización híbrido que se aparte del esquema de vigilancia exclusivamente vertical. En efecto, previas las modificaciones legales necesarias, un sistema que contemple aportaciones de covigilancia provenientes de actores clave (como vecinos de fraccionamientos colindantes, deportistas y visitantes recurrentes) puede potenciar significativamente la incorporación de la tecnología en el proceso de protección del bosque.

La emisión de directrices y lineamientos claros que regulen la aportación de información relevante sobre la posible comisión de delitos o infracciones ambientales por parte de particulares puede potenciar considerablemente la capacidad de actuación del OPD. Bajo este esquema, el organismo deja de ser el único agente de observación sobre el territorio para constituirse en un nodo central de una red de inteligencia colectiva.

Una reforma y reestructuración de esta naturaleza puede aprovechar la presencia constante de los miles de ciudadanos que transitan o visitan el bosque para instituir una función de monitoreo civil activo. Habilitados mediante plataformas digitales oficiales, los ciudadanos acreditados podrían capturar evidencias georreferenciadas (fotografías de descargas ilegales, focos de calor o invasiones al polígono de protección) con valor de prueba documental para el inicio de procedimientos administrativos sancionadores. Esta democratización de la fiscalización no tiene por objeto sustituir a la autoridad, sino dotarla de una visión periférica que elimine las zonas ciegas operativas, permitiendo que los 28 elementos de vigilancia concentren sus esfuerzos de forma estratégica en los puntos de conflicto previamente validados por la red ciudadana.

Más allá de la eficacia operativa, la participación de los ciudadanos en el proceso de fiscalización genera un sentido de corresponsabilidad que desincentiva la comisión de infracciones y delitos ambientales. Cuando la comunidad constata que sus aportaciones tienen un cauce legal definido y que sus reportes activan respuestas de la autoridad en tiempo oportuno, se fortalece tanto el tejido social como la confianza en la institución. Este empoderamiento jurídico de los denominados guardianes ciudadanos crea una barrera social frente a los infractores, quienes dejarían de enfrentarse únicamente a diez guardabosques para confrontar una red de observación conectada y documentada, lo que eleva sustancialmente el costo social y jurídico asociado a la comisión de conductas lesivas para el ecosistema.

Finalmente, la integración de la ciudadanía en la fiscalización digital del bosque ofrece una oportunidad significativa para fortalecer la transparencia y la rendición de cuentas institucional. Un sistema en el que la evidencia es generada por el ciudadano y procesada por la autoridad puede habilitar el seguimiento de cada denuncia hasta su resolución, lo que impondría al OPD la obligación de emitir respuestas concretas y debidamente fundadas. Este cambio de paradigma no solo atiende la brecha operativa entre hectáreas bajo protección y elementos disponibles para vigilarlas, sino que posiciona a Jalisco a la vanguardia de la innovación en gestión pública ambiental, ilustrando que la protección efectiva de la biodiversidad en el siglo XXI depende, en medida determinante, de la capacidad institucional para construir redes de colaboración tecnológica con la propia sociedad civil.

2. El diseño del modelo de fiscalización descentralizada

El modelo propuesto presenta retos y dinámicas aún no suficientemente explorados en el ámbito local de Jalisco, México. Para consolidar un esquema de *smart enforcement*, en los términos propuestos y teorizados por Yadin (2023),

resulta indispensable diseñar con precisión el proceso y atender los aspectos críticos de un modelo híbrido gobierno-ciudadano. El elemento central sobre el que descansa la viabilidad del proceso será, necesariamente, la plataforma y su interfaz. Esta no debe concebirse como un simple aplicativo de recepción de quejas, sino como una terminal de captura de evidencia con valor jurídico, diseñada de forma que el usuario comprenda el alcance de la responsabilidad que asume, las implicaciones que su aportación genera y, concretamente, cómo dicha aportación se traduce en una mejora verificable del proceso de conservación.

En este apartado debe considerarse tanto la información directa como la indirecta que la captura de contenido multimedia por parte del denunciante pueda generar. En efecto, a través de la interfaz, el usuario deberá poder (tomando en cuenta las condiciones propias del entorno de captura, como exposición solar, desplazamiento físico o señal de telecomunicaciones limitada) enviar material que no solo registre gráficamente el evento, sino que incorpore automáticamente metadatos relevantes para el procedimiento de determinación e imposición de sanciones.

Sentado lo anterior, resulta evidente que la protección de la identidad del denunciante mediante un esquema de pseudoanonimato constituye un requisito crítico para promover la adopción masiva del sistema. En un contexto como el nacional, la salvaguarda del denunciante frente a posibles represalias reviste una importancia igual o superior a la calidad e integridad de la información aportada: esta última puede ser filtrada y corroborada por los mecanismos de validación del sistema, mientras que una vulneración a la seguridad del informante produce consecuencias potencialmente irreparables.

Tomando este contexto crítico como punto de partida, el sistema podría generar un identificador único cifrado asociado al dispositivo desde el que se transmite la información; no obstante, la vinculación directa entre ambos elementos podría no ser la solución óptima. Para efectos del sistema de incentivos, el cómputo de participaciones (con independencia de la efectividad procesal de cada denuncia) puede constituir el mecanismo más adecuado, de modo que la identidad civil del denunciante no resulte determinable ni siquiera a partir de las bases de datos en poder de la autoridad.

El sistema deberá contar, cuando menos, con los siguientes elementos: un protocolo de captura asistida de información, orientado a maximizar el valor probatorio del material recabado; un sello de tiempo y ubicación inmutable, mediante cuya generación (a través de un hash digital) se vincule de forma inalterable la hora y el lugar del registro, impidiendo la carga de imágenes de archivo o captadas fuera del polígono de protección; y un módulo de preclasificación que permita al ciudadano identificar, mediante iconografía intuitiva, el tipo de evento observado (tala, incendio, fauna herida, residuos, invasión u otras categorías de incidencia frecuente dentro del ecosistema).

3. El filtrado bajo modelos de aprendizaje profundo

Consolidado el modelo anterior, la probabilidad de recepción masiva de información se torna considerable. Con anterioridad al despliegue de drones de

verificación rápida para corroborar los reportes en campo resulta imperativo clasificar y priorizar, mediante inteligencia artificial, las denuncias que deban recibir mayor peso específico para efectos de la respuesta institucional. La experiencia comparada documenta que uno de los temores fundados de la autoridad al abrir canales de participación ciudadana es precisamente la saturación de los sistemas con reportes irrelevantes, duplicados o deliberadamente malintencionados (Faizah, 2025).

Para atender esta problemática, la incorporación de un filtro de ruido basado en el entrenamiento de redes neuronales (alimentadas con las primeras imágenes validadas o seleccionadas como muestras de referencia de alta calidad) resulta determinante para obtener resultados confiables en el modelo operacional definitivo. Una vez alcanzado un nivel de entrenamiento suficiente, el sistema podrá generar una valoración de veracidad en milisegundos, activando o descartando, según corresponda, la respuesta institucional.

En este paso, se vuelven imperativos algunos elementos para el modelo:

a) Filtrado de duplicados e índice de relevancia por reincidencia: si 50 personas reportan un incidente en un punto cercano, la IA debe poder agrupar y clasificar dicha información para fácil supervisión humana, previniendo una saturación de información de alta relevancia.

b) Análisis de integridad de pixeles: aplicativos de acceso abierto pueden apoyar a detectar si una imagen ha pasado por procesos de edición, generando una categoría de especial revisión humana o que requiera de mayor corroboración para detonar un posible actuar público.

c) Validación geográfica: la IA puede cruzar y referenciar información reportada con datos previos de uso de suelo, detonando la necesidad de acción inmediata en casos de notoria infracción.

d) Determinar grado de acción: todos estos sistemas pueden estar referenciados o cruzados para lograr un índice inequívoco de accionamiento del aparato estatal y sus limitados recursos, tratando de alcanzar el mayor grado de eficiencia operativa posible.

Para esta fase del proceso, resulta indispensable que, con la información previamente clasificada y validada, se integre una matriz de riesgo multidimensional que pondere elementos de alta relevancia, tales como: la proximidad a zonas de alta biodiversidad o de hábitat de fauna sujeta a manejo especial, las condiciones meteorológicas en tiempo real y la recurrencia de incidentes en la zona. Dicha matriz podrá robustecerse de manera progresiva conforme se incremente el volumen de información disponible para el modelo.

Todo lo anterior se sustenta en tecnologías no solo existentes, sino de acceso relativamente asequible, lo que permite que su incorporación no represente un obstáculo insalvable y que pueda alcanzarse una sinergia de alto impacto entre la autoridad y la ciudadanía para efectos de la conservación del patrimonio natural (Loomis *et al*., 2024).

4. Sistema de incentivos

Uno de los mayores desafíos que enfrenta cualquier administración pública es, sin duda, su relación con la población y su capacidad para generar movilización ciudadana en torno a los temas prioritarios. La cuestión no reside en la tecnología propiamente dicha: un modelo que introduce cambios mediante la implementación tecnológica no genera resistencias por esa sola razón; el problema es de naturaleza subyacente y remite a la disposición de las personas para actuar o participar, con independencia del canal o mecanismo a través del cual se les convoque.

La percepción de ineficacia institucional y, en mayor medida aún bajo los supuestos planteados, el temor a represalias, constituyen factores críticos que el diseño del sistema debe atender (Almond y Verba, 1963). Diversos estudios en el ámbito de las ciencias sociales y de ciencia política señalan que la participación ciudadana se incrementa no solo por motivaciones altruistas, sino cuando existe una eficacia percibida genuina: esto es, cuando las acciones de la autoridad se materializan en plazos razonables y sus resultados son comunicados de forma transparente a quienes contribuyeron a generarlos (OECD, 2021).

En el mismo sentido, las teorías de la *gamificación* y del comportamiento prosocial indican que los incentivos con mayor probabilidad de éxito son aquellos que refuerzan la identidad del individuo dentro de su comunidad (Koivisto y Hamari, 2019; Tajfel & Turner, 1986). Conforme a estos planteamientos, las personas participan de forma más activa cuando el sistema ofrece retroalimentación directa y constante sobre el valor de sus aportaciones (Hamari *et al*., 2014). Bajo estos supuestos, un incentivo condicionado a la firmeza de la sanción administrativa contravendría dicha lógica, pues el procedimiento sancionador puede resultar prolongado e impredecible (Hamari *et al*., 2014); en cambio, el modelo propuesto debe articularse en torno a la claridad y consistencia de la colaboración, mediante el establecimiento de una métrica de participación positiva que premie la veracidad y el rigor de la información aportada, con independencia del resultado procesal.

Dicha métrica deberá operar como un sistema de reputación digital en el que, tras filtrar el ruido y el contenido no solicitado (*spam*) mediante algoritmos de inteligencia artificial, el modelo identifique como de mayor valor aquellas denuncias que presenten patrones de evidencia consistentes: geolocalización coherente, imágenes nítidas y coherencia temporal entre el evento reportado y el momento del envío. Las participaciones validadas acumularán puntos con independencia de las acciones que adopte la autoridad, lo que permite reconocer de forma objetiva y continua el valor de la aportación ciudadana como indicador confiable del estado de conservación del bosque.

Este sistema de reputación deberá, a su vez, generar beneficios concretos para los denominados visitantes distinguidos, provenientes tanto de actores estatales involucrados (como los municipios circunvecinos) como de la iniciativa privada. En efecto, el modelo debe contemplar la posibilidad de que cada ciudadano materialice su participación en beneficios tangibles: descuentos en el pago de derechos o contribuciones fiscales, acceso preferente a programas de

desarrollo social, apoyos económicos orientados al fomento del empleo o del emprendimiento, o cualquier otro beneficio que las administraciones públicas competentes determinen como económicamente viable en tanto aportación indirecta a la conservación del Bosque La Primavera.

De forma paralela, la iniciativa privada puede y debe sumarse mediante la suscripción de convenios que permitan que la métrica de participación en la conservación sea canjeable por beneficios en empresas con certificaciones de responsabilidad social. Bajo esta lógica, la participación continua en la protección del bosque genera una ventaja perceptible en la economía cotidiana del ciudadano, configurando un entorno integral en el que la protección del patrimonio natural deja de ser una actividad de retorno exclusivamente social o personal para convertirse en una dinámica de pertenencia y compromiso multidimensional.

A modo de cierre del modelo de incentivos, la articulación directa entre la aportación a la conservación, el reconocimiento del perfil cívico y el beneficio económico tangible reúne las condiciones necesarias para generar un impulso real en la disposición de quienes ya frecuentan ordinariamente el bosque a incorporarse, en la medida de sus posibilidades, al esfuerzo colectivo de conservación. Bajo este supuesto, el ciudadano deja de ser un observador pasivo ante la degradación del Bosque La Primavera para constituirse en un copartícipe activo del fortalecimiento del Estado de derecho, cuyas aportaciones se integran a un sistema que reconoce y premia directamente su rigor, honestidad y compromiso cívico.

5. Consideraciones legales importantes

La eficacia de cualquier sistema de fiscalización ciudadana en entornos de alta vulnerabilidad (donde el infractor puede ser desde un particular descuidado hasta grupos delictivos organizados dedicados a la tala ilegal o a la invasión de predios) depende, en medida determinante, de la seguridad del informante. No obstante, el derecho administrativo sancionador y el sistema de justicia en México imponen que el presunto infractor conozca las pruebas en su contra para poder controvertirlas, en garantía del derecho de audiencia y del debido proceso consagrados en los artículos 14 y 16 de la Constitución Política de los Estados Unidos Mexicanos. De esta conjunción emerge una tensión jurídica de fondo: ¿de qué manera puede una imagen captada por un ciclista en condición de pseudoanonimato sustentar válidamente una sanción administrativa o servir de base para una orden de presentación, sin exponer a su autor a represalias en el marco del procedimiento de ejecución?

La introducción de incentivos para la denuncia ciudadana constituye un instrumento de doble filo. Si bien representa el mecanismo más eficaz para incrementar la cobertura de la fiscalización, su gestión inadecuada puede derivar en una dinámica de obtención indiscriminada de recompensas que sature el sistema con reportes triviales, malintencionados o deliberadamente fabricados para obtener beneficios económicos o para perjudicar a terceros, como podría ocurrir entre desarrolladores inmobiliarios o propietarios de predios colindantes. El diseño ético del sistema exige, por tanto, que el eje gravitacional del modelo

transite del interés pecuniario hacia el deber cívico, sin desconocer por ello la legitimidad del reconocimiento al esfuerzo que entraña el ejercicio sostenido de la vigilancia ciudadana.

El régimen sancionador aplicable debe establecer con precisión que el uso de la plataforma para fines de acoso, denuncia falsa o fraude constituye una infracción administrativa grave, susceptible de generar responsabilidad tanto en la vía administrativa como, en su caso, en la penal. Esta tipificación genera por sí misma un desincentivo estructural frente a la obtención indiscriminada de recompensas. A ello se suman, como primer filtro de naturaleza técnico-ética, los mecanismos de validación mediante visión computacional previamente descritos. Ante un intento de fabricación de evidencia (como el traslado de residuos a un predio ajeno con el propósito de denunciarlo), el sistema de inteligencia artificial, al analizar los metadatos y la secuencia temporal del reporte, se encuentra en condiciones de detectar patrones de comportamiento incompatibles con una observación espontánea y contextualmente coherente, activando los protocolos de revisión humana que correspondan.

III. Conclusiones

El estado actual que guarda la conservación del área natural protegida Bosque La Primavera, con las limitantes ya establecidas, no es una falla de voluntad en estricto sentido, claramente es un síntoma del agotamiento del modelo de vigilancia vertical que impera en México y en diversas partes del mundo. Como quedó claramente establecido, la extrema disparidad entre las hectáreas protegidas y los recursos humanos disponibles para ello crea un escenario sumamente desfavorable para el OPD pero sobre todo para el bosque mismo. La incorporación de la tecnología, pero sobre todo de las personas en el proceso no es una idea caprichosa de modernización, es una necesidad imperativa y una propuesta real, aterrizada y alcanzable para rescatar los *parques de papel* y transformarlos en ecosistemas vivos, participativos con protección material efectiva.

El modelo de fiscalización descentralizada propuesto rompe con el monopolio de observación estatal, integrando a las personas no como simples denunciantes sino como generadores del proceso de procesamiento inteligente de información masiva. Al tomar como base los casos de éxito como el sistema tributario mexicano o el gobierno digital de Estonia, es evidente que la tecnología aplicada de forma adecuada tiene la capacidad de aumentar de forma exponencial el alcance de la autoridad sin necesidad de un aumento imposible de presupuesto público. La implementación de una arquitectura de sensores humanos validados por inteligencia artificial permite que el Organismo Público Descentralizado sortee sus limitaciones físicas, convirtiendo cada dispositivo móvil en un coadyuvante y por lo tanto extensión de la acción pública.

Un aspecto crítico de la presente propuesta es que la tecnología, por sí sola, no alcanzará la meta planteada: la aportación y la voluntad humanas constituyen el factor determinante del éxito del modelo. En este sentido, el sistema de incentivos articulado en torno a la métrica de participación positiva resulta fundamental para el establecimiento de un tejido social colectivo orientado a la

protección del bosque. Al vincular el civismo ambiental con beneficios tangibles, como estímulos fiscales o acceso preferente a programas de desarrollo social, el modelo alinea el interés individual con el bienestar ecológico, configurando un entorno en el que la protección del patrimonio natural se convierte en una actividad económicamente racional y socialmente reconocida.

No obstante, la viabilidad de esta propuesta depende estrictamente de la solidez técnica del filtrado de datos. La incorporación de modelos de aprendizaje profundo para detectar duplicados, validar la integridad de los píxeles y analizar metadatos geográficos es el «primer filtro ético» que previene la perversión del sistema. Esta capa de inteligencia artificial no solo optimiza la operatividad de los escasos cuerpos de guardabosques, sino que blinda el proceso sancionador contra intentos de fraude o «caza de recompensas», garantizando que el aparato estatal solo se active ante incidentes de alta probabilidad y relevancia.

Desde la perspectiva jurídica, el desafío del pseudoanonimato representa la frontera final para la democratización de la justicia ambiental en contextos de alta vulnerabilidad. El diseño de identificadores cifrados vinculados a dispositivos permite proteger la integridad del ciudadano frente a posibles represalias de grupos delictivos, manteniendo al mismo tiempo la trazabilidad necesaria para el debido proceso administrativo. Este equilibrio entre seguridad y legalidad es lo que permitirá que la información recolectada por un senderista se convierta en una prueba documental inexpugnable, cerrando la brecha histórica entre el daño detectado y la sanción ejecutada.

Finalmente, el presente estudio concluye que el futuro de la conservación en México depende, en medida determinante, de la capacidad institucional para construir redes de colaboración tecnológica con la propia sociedad civil. La protección del Bosque La Primavera no puede continuar recayendo sobre un cuerpo de diez guardabosques; debe ser asumida como responsabilidad compartida de una red ciudadana conectada, vigilante y habilitada por herramientas tecnológicas de cuarta generación. Al adoptar este modelo de fiscalización descentralizada, Jalisco y México tienen la oportunidad de encabezar un cambio de paradigma con proyección global, en el que el Estado no solo custodie el territorio, sino que gestione la inteligencia colectiva como recurso institucional para garantizar que la justicia ambiental transite, de forma definitiva, de un modelo reactivo a uno verdaderamente preventivo y eficaz.

Referencias bibliográficas

Almond, G. A., Verba, S. (1963). *The civic culture: Political attitudes and democracy in five nations*. Princeton University Press.

Carpio-Domínguez, J. L. (2024). Aplicación de la ley forestal frente al cambio climático. Un análisis criminológico verde en el contexto mexicano (periodo 2009-2020). *REC: Revista Electrónica de Criminología*, *4*. https://revistaseug.ugr.es/index.php/REC/article/view/33164

Castro Salazar, J. I., Carpio Domínguez, J. L. y Arroyo Quiroz, I. (2021). Acciones y limitantes institucionales en la aplicación de la legislación forestal en

México en el periodo 2009-2019. *Revista de El Colegio de San Luis, 11*(22), e1325. https://doi.org/10.21696/rcsl112220211325

Centro de Estudios y Tecnología del Sureste (CETYS). (2019). *La transformación digital en los procesos de fiscalización*. CETYS Universidad. https://www.cetys.mx/noticias/la-transformacion-digital-en-los-procesos-de-fiscalizacion

Cervantes, E.; Martínez, J. y Núñez, L. (2023). *Matar el futuro: La 4T y el fin de la política ambiental en México*. Mexicanos contra la Corrupción y la Impunidad. https://contralacorrupcion.mx/matar-el-futuro-la-4t-y-el-fin-de-la-politica-ambiental-en-mexico

City of Amsterdam. (2024). *Policy: Innovation*. Government of the City of Amsterdam. https://www.amsterdam.nl/en/policy/policy-innovation/

City of Santa Cruz. (2024). *Artificial Intelligence (AI) – AI Transparency Portal*. City of Santa Cruz Government. https://www.santacruzca.gov/Government/City-Departments/Information-Technology/Artificial-Intelligence-AI

Comisión Nacional Forestal (Conafor). (2024). *Reporte semanal de incendios forestales*. Gobierno de México. https://www.gob.mx/conafor/documentos/reporte-semanal-de-incendios

Criado, J. I. (2021). Inteligencia Artificial (y Administración Pública). *Eunomía. Revista en Cultura de la Legalidad,* (20), 348-372. https://doi.org/10.20318/eunomia.2021.6097

Facturando.mx. (2026, 10 de abril). *SAT 2025: Eficiencia en recaudación y fiscalización México*. https://www.facturando.mx/blog/index.php/2026/04/10/recaudacion-sat-2025-eficiencia

González Cuevas, G. A. (2023, 3 de diciembre). Declaraciones sobre incendios forestales en Jalisco [Entrevista a investigador del CUCBA-UdeG]. *El Informador*. https://www.informador.mx/amp/jalisco/Incendios-forestales-El-2023-ano-con-mas-incendios-En-Bosques-20231203-0022.Html

Government Digital Service; Office for Artificial Intelligence; Department for Science, Innovation and Technology. (2019, 10 de junio). Using data from electricity meters to predict energy consumption. *Gov.uk*. https://www.gov.uk/government/case-studies/using-data-from-electricity-meters-to-predict-energy-consumption

Hamari, J., Koivisto, J. y Sarsa, H. (2014). Does gamification work? A literature review of empirical studies on gamification. En *Proceedings of the 47th Hawaii International Conference on System Sciences* (pp. 3025-3034). IEEE. https://doi.org/10.1109/HICSS.2014.377

Huerta-Martínez, F. M. e Ibarra-Montoya, J. L. (2014). Incendios en el bosque la primavera (Jalisco, México): un acercamiento a sus posibles causas y consecuencias. *CienciaUAT, 9*(1), 91-105. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-78582014000100023&lng=es&tlng=es

Jardel, E. (2023, 30 de mayo). Declaraciones sobre incendios forestales en Jalisco [Entrevista a investigador del Departamento de Ecología y Recursos Naturales-UdeG]. *Mongabay en Español*. https://es.mongabay.com/2023/05/jalisco-perdio-100-mil-hectareas-de-bosques-incendios-mexico

Koivisto, J. y Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management, 45,* 191-210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

LexLatin. (2026, 29 de enero). El SAT refuerza fiscalización con tecnología: su Plan Maestro y las auditorías digitales. *LexLatin*. https://lexlatin.com/reportajes/sat-fiscalizacion-tecnologia-plan-maestro-auditorias-digitales

Machlis, G. E. y Tichnell, D. L. (1985). *The State of the World's Parks: An International Assessment for Resource Management, Policy, and Research*. Westview Press.

OCDE y CAF. (2023). *Revisión del gobierno digital OCDE-CAF en América Latina y el Caribe: Construyendo servicios públicos inclusivos y responsivos*. OCDE Publishing. https://www.caf.com/es/actualidad/noticias/lecciones-de-la-ocde-y-caf-sobre-gobierno-digital-en-latinoamerica

OECD Observatory of Public Sector Innovation. (2024a). *AI and geospatial data to protect against wildfires*. OECD OPSI Case Study Library. https://oecd-opsi.org/innovations/ai-and-geospatial-data-to-protect-against-wildfires/

OECD Observatory of Public Sector Innovation. (2024b). *Enhancing urban resilience with artificial intelligence and digital twins*. OECD OPSI Case Study Library. https://oecd-opsi.org/innovations/enhancing-urban-resilience-with-artificial-intelligence-and-digital-twins/

OECD Observatory of Public Sector Innovation. (2024c). *Establishment of an AI Competence Centre for the City of Munich*. OECD OPSI Case Study Library. https://oecd-opsi.org/innovations/establishment-of-an-ai-competence-centre-for-the-city-of-munich/

OECD Observatory of Public Sector Innovation. (2024d, July 1). *From Innovation Challenge to AI-powered waste diversion*. OECD OPSI Case Study Library. https://oecd-opsi.org/innovations/ai-for-saskatchewans-waste-diversion/

OECD Observatory of Public Sector Innovation. (2024e). *vCity, a human-centric platform for urban digital twins*. OECD OPSI Case Study Library. https://oecd-opsi.org/INNOVATIONS/VCITY-A-HUMAN-CENTRIC-PLATFORM-FOR-URBAN-DIGITAL-TWINS/

OECD. (2021). *Political efficacy and participation*. OECD Publishing. https://doi.org/10.1787/4548cad8-en

Organismo Público Descentralizado Bosque La Primavera. (2023). *Reporte de incendios forestales 2023*. Gobierno del Estado de Jalisco.

Organismo Público Descentralizado Bosque La Primavera. (2024). *Plan Institucional del Organismo Público Descentralizado Bosque La Primavera*. Gobierno del Estado de Jalisco. https://plan.jalisco.gob.mx/wp-content/uploads/2024/11/Organismo-Publico-Descentralizado-Bosque-La-Primavera.pdf

Procuraduría Federal de Protección al Ambiente (Profepa). (s. f.). *Universo de atención en materia forestal*. http://www.profepa.gob.mx/innovaportal/v/255/1/mx/universo_de_atencion_en_materia_forestal.html

Seoul Metropolitan Government. (2024). *Big Data & AI – Big Data Service Platform*. Government of Seoul. https://english.seoul.go.kr/policy/smart-city/big-data-ai/

Servicio de Administración Tributaria (SAT). (2014). *Contabilidad electrónica*. Gobierno de México. http://omawww.sat.gob.mx/fichas_tematicas/reforma_fiscal/Paginas/contabilidad_electronica.aspx

Singapore Land Authority. (2020). *Virtual Singapore – Singapore's virtual twin*. Government of Singapore. https://www.sla.gov.sg

Tajfel, H. y Turner, J. C. (1986). The social identity theory of intergroup behavior. En S. Worchel y W. G. Austin (Eds.), *Psychology of intergroup relations* (2.ª ed., pp. 7-24). Nelson-Hall.

Valdez-Rosas, J. S. (2023). Revisión de causas, consecuencias y medidas de respuesta frente a los incendios forestales: un enfoque en el estado de Jalisco. *e-CUCBA, 20*(19). http://e-cucba.cucba.udg.mx/index.php/e-Cucba/article/view/327

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 206-223

JUSTICIA JUVENIL ALGORÍTMICA: DESAFÍOS DE LA EVIDENCIA ACTUARIAL Y EL DERECHO A LA DEFENSA EN EL SISTEMA DE RESPONSABILIDAD PENAL ADOLESCENTE

ALGORITHMIC JUVENILE JUSTICE: CHALLENGES POSED BY ACTUARIAL EVIDENCE AND THE RIGHT TO A DEFENSE IN THE JUVENILE CRIMINAL JUSTICE SYSTEM

Pedro Luis Bracho-Fuenmayor
Universidad Tecnológica Metropolitana, Chile

Resumen

El artículo examina la compatibilidad existente entre evidencia actuarial y el sistema chileno de responsabilidad penal adolescente, particularmente, a partir de la incorporación del sistema de evaluación y toma de decisiones (SEYTD). En tal sentido, el trabajo se enmarca en el paradigma de una investigación jurídica dogmática con enfoque analítico crítico, en la cual se examina el marco normativo de la especialidad, la individualización de la sanción y el interés superior del adolescente, a la vez que analiza el estatuto jurídico de los instrumentos estructurados de evaluación del riesgo. En efecto, se sostiene que la evidencia actuarial derivada no es ilegítima *per se* pero puede devenir en problemática cuando sus resultados adquieren un valor autónomo capaz de desplazar el juicio individualizador y por ende debilitando el derecho a la defensa. Sobre esa base, el trabajo muestra que la dificultad central no se encuentra radicada en la mera existencia de estos instrumentos, sino en las condiciones bajo las cuales sus resultados pretenden ingresar con relevancia decisional al sistema penal adolescente. Consecuentemente, se concluye que la técnica solo es admisible si permanece jurídicamente subordinada a la especialidad y singularidad del caso.

Palabras clave

Justicia penal juvenil, legitimidad, interés superior, garantías, contradicción

Abstract

The article examines the compatibility between actuarial evidence and the Chilean system of criminal responsibility for adolescents, particularly following the introduction of the Assessment and Decision-Making System (SEYTD). In this regard, the study is framed within the paradigm of dogmatic legal research with a critical analytical approach, in which the regulatory framework of the field, the individualization of punishment and the best interests of the adolescents are examined, whilst analyzing the legal status of structured risk assessment tools. Indeed, it is argued that the derived actuarial evidence is not illegitimate per se but can become problematic when its results acquire an autonomous value capable of supplanting the individualizing judgement and thereby undermining the right to defense. On this basis, the paper demonstrates that the central difficulty does not lie in the mere existence of these instruments, but in the conditions under which their results are intended to enter the juvenile criminal justice system with decision-making relevance. Consequently, it is concluded that the technique is only admissible if it remains legally subordinate to the specific nature and uniqueness of the case.

Keywords

Juvenile criminal justice, legitimacy, best interest, guarantees, contradiction

I. Introducción

La justicia penal juvenil contemporánea se encuentra atravesada por una transformación de particular relevancia jurídica (Sánchez y Lescano-Galeas, 2021), la cual plantea el desplazamiento desde modelos centrados en la apreciación individual del caso hacia esquemas de evaluación estructurada del riesgo (Alarcón *et al*., 2022; Kim *et al*., 2019); tal giro actuarial no supone solo la incorporación de herramientas técnicas, sino una modificación en la racionalidad de la intervención estatal (Avello Saez *et al*., 2018).

De allí que, donde antes se planteaba el predominio la reconstrucción cualitativa de la trayectoria del adolescente, hoy adquieren creciente protagonismo las lógicas de clasificación, puntuación y gestión de probabilidades. En este sentido, la literatura especializada ha mostrado, como se verá *ut infra,* que tales instrumentos aumentan la consistencia decisional, a la vez que advierten que esa pretensión de consistencia no resuelve, por sí sola, los problemas normativos que ven la luz cuando dichas herramientas ingresan al ámbito de las garantías procesales y los derechos fundamentales.

En Chile, esta discusión ha dejado de ubicarse en el plano puramente teórico, toda vez que la Ley 21.527 dio vida al Servicio Nacional de Reinserción Social Juvenil como entidad especializada para la administración y ejecución de las medidas y sanciones de la Ley 20.084, a la vez que ordena la implementación de un modelo de intervención especializado de aplicación nacional y vinculante.

Consecuentemente, en ese marco se inserta el Sistema de Evaluación y Toma de Decisiones (SEYTD), el cual se presenta en la normativa técnica como una herramienta transversal para la evaluación, planificación y seguimiento de la intervención; de esta manera, el sistema chileno se incorpora a la tendencia comparada hacia el uso de la evidencia actuarial en justicia juvenil.

Sin embargo, tal reconfiguración institucional muestra un límite normativo especialmente exigente, en razón de que el régimen penal adolescente chileno se encuentra planteado como un sistema diferenciado (Bracho-Fuenmayor & Pérez Cobo, 2025), el cual se orienta por el principio de especialidad, el interés superior y la individualización de la respuesta penal.

De allí que la Ley 20.084 fije una finalidad socioeducativa orientada a la plena integración social y exige que la determinación de la sanción considere la edad, el desarrollo psicosocial y los antecedentes del caso; a ello se suma el artículo 40 de la Convención sobre los Derechos del Niño (1990), el cual impone un trato compatible con la dignidad del adolescente, su edad y su reintegración, así como la Observación General n.º 24 del Comité de los Derechos del Niño, el cual enfatiza el carácter especializado del sistema de justicia juvenil.

En este contexto la legitimidad de la intervención no depende solo de su eficacia administrativa, sino de su compatibilidad con un tratamiento individualizado y respetuoso de la posición jurídica reforzada del adolescente. Precisamente, de allí emerge el problema que vertebra esta investigación, debido a que mientras la individualización exige una comprensión situada del caso concreto, la evidencia actuarial opera sobre regularidades poblacionales, correlaciones y perfiles de

riesgo que no equivalen en modo alguno, a una explicación singular de la conducta ni a una justificación suficiente de la respuesta penal en un caso determinado.

La dificultad se vuelve aún más patente, porque esta evidencia no opera como antecedente auxiliar de escasa relevancia, sino como un soporte técnico con capacidad real que incide en la intensidad de la intervención prevista en el plan individual del adolescente; en ese contexto la cuestión decisiva ya no versa sobre si el modelo matriz de intervención diferenciada con adolescentes ofrece algún rendimiento predictivo, sino si su funcionamiento puede ser jurídicamente controlado.

Así, cuando el puntaje o la categoría de riesgo se presenta sin información suficiente sobre las variables consideradas, las reglas de puntuación, los procesos de validación, sus límites o márgenes de error, la defensa puede quedar en una posición de desventaja técnica frente a una conclusión emanada del informe de evaluación integral que le afecta de manera directa, pero cuyo proceso de construcción no logra reconstruir ni controvertir de forma eficaz.

A ello, ha de añadirse un reparo adicional, que es el riesgo de que variables atravesadas por desigualdad estructural operen como *proxies* de vulnerabilidad social y proyecten sobre el adolescente una peligrosidad que refleje, más bien, trayectorias de exclusión o patrones de sobreintervención institucional.

De allí que el problema no consiste en sostener que toda herramienta actuarial sea intrínsecamente ilegítima o necesariamente discriminatoria, sino en advertir que, sin transparencia accionable, auditoría y control contradictorio suficientes, el sistema carece de resguardos idóneos para detectar y corregir impactos indebidos especialmente en una población cuya protección jurídica debe ser reforzada.

En tal sentido, la utilización de evidencia actuarial derivada del SEYTD así como de instrumentos asociados, no es, por sí misma, jurídicamente ilegítima; sin embargo, esta puede volverse incompatible con las exigencias de especialidad (Carrasco Jiménez, 2019), individualización y derecho a la defensa cuando sus resultados se emplean sin un estándar suficiente de fiabilidad, transparencia y contradicción, adquiriendo además un peso tal que terminan desplazando el juicio interdisciplinario e individualizador que debe orientar a la justicia penal adolescente.

Desde esa premisa, se examina, el estatuto normativo que dota de especialidad y da individualización en el sistema de responsabilidad penal de los adolescentes en Chile; es así como, se deja ver la configuración institucional de la evidencia actuarial en la arquitectura del sistema de evaluación y toma de decisiones (SEYTD); consecuentemente se propone un estándar de compatibilidad dogmática orientado a someter su eventual utilización a condiciones reforzadas de control procesal.

Este trabajo se desarrolla a través de una investigación jurídico-dogmática de orientación analítico-crítica dirigida a determinar si la incorporación de evidencia actuarial en el sistema chileno de responsabilidad penal adolescente (Graepp y Ramírez, 2003), resulta compatible con las exigencias de especialidad, individualización y derecho a la defensa.

El estudio se articula con la interpretación sistemática del marco normativo aplicable, en particular las leyes 20.084, 21.527 y 21.430, junto con la Convención Sobre los Derechos del Niño y sus estándares de interpretación, aunado a la normativa técnica e institucional relativa al modelo de intervención especializado, así como al sistema de evaluación y toma de decisiones.

A ello, se suma el contraste con doctrina especializada, aunado a las experiencias comparadas relevantes en materia de opacidad técnica, prueba de fiabilidad y condiciones de contradicción necesarias en los instrumentos estructurados de evaluación del riesgo (Bracho-Fuenmayor, 2025a), destacando que el análisis no persigue medir empíricamente la frecuencia de uso ni resultados judiciales, sino reconstruir el estatuto jurídico de esta evidencia técnica, identificar sus posibles puntos de fricción con el sistema de garantías y formular un estándar de compatibilidad dogmática para su eventual utilización bajo condiciones reforzadas de trasparencia, control contradictorio y subordinación a la individualización propia de la justicia penal juvenil.

II. Especialidad, individualización y justicia penal adolescente

La justicia penal adolescente no constituye en modo alguno una versión simplificada y atenuada del sistema penal de adultos, ni mucho menos una mera adaptación procedimental del Derecho penal común; toda vez, que se trata de un régimen jurídicamente diferenciado, construido sobre la base de la especialidad del sujeto, de la especificidad de la intervención estatal y de la finalidad propia que ha de cumplir la respuesta sancionatoria frente a los adolescentes en conflicto con la ley penal. Por su parte, la Ley 20.084 regula la responsabilidad penal de los adolescentes, las consecuencias jurídicas aplicables y su ejecución, mientras que la Ley 21.527 crea una institucionalidad especializada para la administración y ejecución de tales medidas y sanciones (Ibarra y Barragán, 2007).

De allí que, la especialidad no puede ser entendida en un sentido puramente orgánico o administrativo; toda vez que su alcance es más exigente en la medida que expresa la improcedencia de tratar al adolescente infractor como si se encontrara en la misma posición normativa (González Laurino *et al*., 2024; Ramírez, 2007), biográfica y evolutiva que un adulto penalmente responsable. Ahora bien, la existencia de un sistema diferenciado responde a la necesidad de que la decisión penal se construya desde la condición del sujeto en desarrollo progresivo y evolutivo del adolescente y no desde una lógica indiferenciada de imputación, sanción o control.

Esta comprensión se encuentra reforzada por el Derecho internacional de los derechos humanos, en el artículo 40 de la Convención sobre los Derecho del Niño en la cual se exige que todo niño acusado o declarado responsable de infringir la ley penal sea tratado de un modo compatible con su dignidad y valor, teniendo en cuenta su edad y la finalidad de promover su reintegración y el desempeño de un papel constructivo en la sociedad.

A esto se añade la observación General n.º 24 del Comité de los Derechos del Niño, que insiste en el carácter especializado del sistema de justicia juvenil

y en la necesidad de asegurar garantías procesales plenas dentro de un modelo compatible con la dignidad, participación y la reintegración.

Desde esta perspectiva, la especialidad no constituye un rasgo accesorio del sistema, sino un verdadera exigencia de legitimidad de toda decisión que incida sobre la libertad, la sanción o el itinerario de reinserción del adolescente (Argudo *et al*., 2021), por ello la respuesta estatal solo puede reputarse ajustada a derecho si resulta coherente con la posición jurídica reforzada que el ordenamiento reconoce al adolescente como sujeto de derechos.

La proyección más concreta de esta especialidad se ve proyectada en la categoría del interés superior y la individualización de la respuesta penal; en ese sentido la Ley 21.430 reconoce el interés superior como derecho (Berríos *et al*., 2025), principio y norma de procedimiento mientras que la Ley 20.084 y la Ley 21.527 ordenan tenerlo especialmente presente en todas las actuaciones judiciales y administrativas relativas a adolescentes sometidos al sistema penal juvenil (Vera Vega y Mayer Lux, 2024).

En este ámbito, el interés superior del adolescente en modo alguno, autoriza prácticas paternalistas que devengan en márgenes de discrecionalidad inmotivadas, ello en razón de que su función es orientar la intervención estatal hacia la mayor satisfacción posible de sus derechos dentro de un sistema de responsabilidad (Berríos Díaz, 2011a); desde esa lógica, la individualización no puede entenderse como una técnica de dosificación de la sanción, sino como una exigencia normativa que deriva directamente del principio de especialidad.

En respaldo de lo esgrimido *ut supra*, la Ley 20.084 fija una finalidad socioeducativa orientada a la plena integración social del adolescente, exigiendo desarrollo psicosocial y los antecedentes del caso. De ello se deriva que la legitimidad de la sanción depende de su construcción a partir de una comprensión situada del adolescente, de sus condiciones de desarrollo y del sentido reintegrador que debe asumir la intervención del Estado.

Precisamente en este punto emerge la tensión central con la lógica actuarial, dado que la individualización exige justificar una decisión respecto de un sujeto singular, situado en una trayectoria vital concreta y dotada de un potencial de cambio que no puede ser reducido a categorías abstractas ni meras expectativas, sino de una manera reforzada de acuerdo con las respectivas especificidades que la especialidad exige y las condiciones propias, ameritan.

La evidencia actuarial en cambio, opera a partir de regularidades poblacionales, correlaciones y perfiles de riesgo que permiten clasificar casos según la probabilidad de determinados desenlaces, a tal efecto, entre ambos planos no existe una relación de identidad conceptual, pues, que un instrumento arroje una categoría de riesgo puede constituir un antecedente técnico relevante, pero no equivale a una explicación suficiente del caso ni reemplaza la fundamentación jurídica exigible en un sistema de especialidad reforzada.

La dificultad entonces, no radica en la mera incorporación de información estructurada, sino en el estatuto que esa información asume dentro de la decisión penal juvenil, si el dato actuarial funciona como insumo subordinado a una valoración judicial e interdisciplinaria auténticamente individualizadora, su

presencia puede resultar compatible con el sistema; pero si la clasificación o la recomendación técnica adquieren un peso capaz de desplazar la reconstrucción singular del caso y de operar como atajo cognitivo para la decisión entonces se resiente el núcleo mismo de la especialidad.

Desde esta perspectiva, la especialidad no solo describe la fisonomía general del sistema penal adolescente, sino que actúa como una suerte de criterio de admisibilidad y límite de relevancia jurídica de cara a la tecnificación de la respuesta penal. En ese orden de ideas, la Ley 21.527 promueve un modelo de intervención especializado, basado en evidencia, pero ello en modo alguno autoriza a entender que la consistencia técnica o la eficiencia administrativa basten para legitimar una decisión que incide sobre derechos fundamentales.

En justicia penal adolescente, la tecnología solo puede ser admitida si permanece subordinada a una exigencia reforzada de comprensión, control y fundamentación, en consecuencia, ninguna clasificación técnica por más estandarizada que sea, puede transformarse en razón autosuficiente para definir la intensidad de la intervención ni para orientar de manera determinante la respuesta institucional.

La especificidad del sistema impide aceptar que la gestión del riesgo sustituya la lógica jurídica de la individualización, de ahí que, antes de examinar el diseño, los instrumentos de rendimiento del SEYTD, resulta indispensable fijar este parámetro dogmático, en el que toda evidencia actuarial que pretenda ingresar legítimamente al sistema penal adolescente debe hacerlo bajo condiciones que aseguren su subordinación al juicio individualizador, al interés superior y a la estructura de garantías propia de una justicia penal juvenil especializada.

III. Evidencia actuarial y configuración del SEYTD en Chile

La configuración chilena de la evidencia actuarial en justicia penal adolescente debe situarse en la nueva arquitectura institucional devenida a raíz de la Ley 21.527, la cual crea el Servicio Nacional de Reinserción Social Juvenil, como servicio público descentralizado, bajo la vigilancia del Ministerio de Justicia y Derecho Humanos, definiéndola como la entidad especializada responsable de administrar y ejecutar las medidas y sanciones contempladas por la Ley N° 20.084 garantizando además el pleno respeto de los derechos humanos de sus sujetos de atención.

Se destaca que, la referida ley somete al servicio al principio de especialización y orienta su gestión hacia la integración social de quienes quedan sujetos a las medidas y sanciones del sistema penal adolescente (Torres-Vásquez y Tirado-Acero, 2023). En ese nuevo marco, la reforma no solo introdujo una institucionalidad distinta del antiguo Sename, sino también una racionalidad de intervención apoyada en herramientas técnicas de evaluación y planificación más coherentes con un modelo de gestión especializado y de aplicación nacional.

En concordancia con el desarrollo de ese marco, la normativa técnica identifica como ejes centrales de la reforma el modelo de intervención especializado y el sistema de evaluación y toma de decisiones (SEYTD); así pues, en esa misma

línea, las normas técnicas aprobadas para la ejecución de sanciones y programas remiten expresamente tanto a la resolución que aprobó dicho modelo como a aquella que estableció el SEYTD, al que reconocen como el instrumento que estructura la evaluación de la intervención.

En ese sentido, el SEYTD no se configura normativamente como un sistema autónomo ni como una herramienta propietaria de caja negra, sino como una arquitectura pública y reglada de gestión de casos, la cual se ve integrada al diseño general del nuevo servicio; su relevancia jurídica radica en que estructura la forma en que la información del caso se procesa y se proyecta en decisiones de intervención, de allí que la dimensión propiamente actuarial del sistema se haga visible cuando la normativa operacionaliza la determinación del riesgo de reincidencia y la vincula con la intensidad de la intervención.

En las normas técnicas aplicables a programas de medio libre, como la libertad asistida simple y la libertad asistida especial, se dispone que la o él profesional gestor de casos debe aplicar instrumentos consignados por el SEYTD para determinar factores y nivel de riesgo de reincidencia (Zambrano *et al*., 2023), la misma normativa establece que los niveles de riesgo pueden ser bajo, moderado o alto y dispone de forma expresa que a mayor riesgo de reincidencia la intervención deberá ser más intensa, de este modo, la evidencia actuarial no permanece en el nivel de una simple descripción técnica, sino que adquiere relevancia práctica en la definición del tipo, alcance e intensidad del proceso de intervención.

Esta lógica se complementa con una distinción técnicamente significativa entre factores de tipos estáticos y dinámicos o necesidades criminógenas. La normativa considera, por una parte, factores estáticos asociados tanto con la trayectoria delictiva, como con el historial previo del adolescente y por otra, con factores dinámicos vinculados con dimensiones susceptibles de intervención, tales como actitudes pro criminales, soportes sociales del delito (Basualto, 2007), relaciones familiares, inserción escolar o laboral, uso del tiempo libre y consumo de sustancias.

En este sentido, la relevancia de esta diferenciación radica en que mientras que el sistema no solo clasifica el riesgo sino que también organiza la intervención a partir de variables que pretenden orientar el trabajo especializado sobre condiciones que podrían modificarse en el curso del proceso (Salas, 2012), en consecuencia, el SEYTD no debe ser entendido como un simple mecanismo de puntuación, sino como una estructura técnico-institucional que articula evaluación, categorización y planificación.

En cuanto a los instrumentos específicos del formulario de evaluación de riesgos y reincidencia (FER-R) cuenta con un respaldo empírico nacional verificable, toda vez que se examinó su validez predictiva en adolescentes infractores de ley (Mettifogo *et al*., 2015; Ramírez, 2017), mostrando que se trata de una herramienta estructurada orientada a registrar dominios de riesgos criminógenos, recursos e indicadores vinculados con la trayectoria delictiva.

Tal constatación resulta relevante porque permite afirmar que al menos respecto de uno de los instrumentos empleados por el sistema, existe una base técnica susceptible de discusión académica y no una mera práctica administrativa

desprovista de soporte verificable (Malacarne y De Azevedo, 2022); sin embargo, dado que estos instrumentos producen resultados con incidencia potencial en la intensidad de la intervención, es por lo que precisamente, su estatuto jurídico, no puede agotarse en la sola constatación de su utilidad técnica o de su rendimiento predictivo.

Desde una perspectiva dogmática, la consecuencia principal es que la evidencia actuarial relevante en el sistema chileno no se materializa en un algoritmo opaco en sentido fuerte, sino en una cadena institucionalmente estructurada de instrumentos, categorías de riesgo, hipótesis de intervención y planes individuales de trabajo.

Tal configuración no disminuye su importancia jurídica, contrariamente, la intensifica, toda vez que se trata de una arquitectura pública, normativamente reglada y con incidencia efectiva en la definición de la intervención, de ahí que el problema central no radique en su mera existencia, sino en el estatuto que sus resultados pueden asumir dentro de un sistema que sigue regido por la especialidad, la individualización y el derecho a la defensa. Sobre esa base, el examen del SEYTD no puede detenerse en su descripción institucional, sino que debe avanzar hacia la cuestión decisiva de sus límites dogmáticos y procesales.

IV. Límites dogmáticos de la evidencia actuarial

La primera limitación dogmática de la evidencia actuarial en justicia penal adolescente reside en que la lógica de predicción del riesgo no coincide con la lógica de justificación de la respuesta penal. En tal sentido, los instrumentos actuariales operan sobre correlaciones, frecuencias y perfiles agregados; su rendimiento depende de la capacidad de clasificar casos según la probabilidad de determinados desenlaces.

La decisión penal juvenil, en cambio, exige una fundamentación construida respecto de un sujeto singular, situado en una trayectoria vital concreta y amparado por un estatuto reforzado de derechos; de ahí que la categoría de riesgo, aun cuando pueda constituir un antecedente técnico relevante para ordenar la intervención, no equivalga por si misma a una razón jurídica suficiente para determinar la intensidad de la respuesta estatal.

En consecuencia, confundir ambos planos supone desplazar la individualización exigida por el sistema hacia una racionalidad de gestión de probabilidades que por definición trabaja con regularidades poblacionales y no con la singularidad del caso; destacando que tal tensión aparece claramente, lo cual deja ver que, aunque útil para estructurar la intervención resulta insuficiente para reemplazar la justificación individualizadora que demanda la justicia juvenil.

Desde esta perspectiva el problema no consiste en negar de modo abstracto toda relevancia a la información estructurada, sino en fijar su estatuto dogmático dentro del sistema, en un régimen orientado por la especificidad, el interés superior y la finalidad socioeducativa de la sanción (Velásquez *et al*., 2023), el riesgo no puede transformarse en criterio autónomo de decisión ni en sucedáneo del juicio interdisciplinario.

La evidencia actuarial solo resulta compatible con ese marco si opera como insumo subordinado a una valoración jurídica e individualizadora genuina, cuando por el contrario la clasificación o la recomendación técnica adquieren un peso capaz de funcionar como atajo cognitivo para definir la respuesta institucional, se produce una inversión dogmáticamente problemática, pues el instrumento deja de asistir a la decisión y comienza a gobernarla; en ese punto, la especialidad deja de actuar como parámetro rector y la intervención corre el riesgo de configurarse en clave de administración de peligrosidad, desplazando el núcleo personal, contextual y evolutivo que la justicia penal adolescente está obligada a preservar.

A ello, se suma un segundo límite, vinculado con el modo en que la evidencia actuarial trata la relación entre trayectoria delictiva, condiciones sociales y pronóstico institucional. El diseño normativo del SEYTD distingue factores estáticos de factores dinámicos o necesidades criminógenas, y esa distinción es técnicamente relevante porque permite estructurar planes de intervención orientados al cambio.

Sin embargo, desde el punto de vista dogmático, la incorporación de variables estrechamente atravesadas por desigualdad estructural obliga a una cautela reforzada; en este sentido dominios tales como desvinculación escolar, supervisión familiar débil, redes de pares desadaptados o contextos comunitarios precarios pueden funcionar, al menos potencialmente como *proxies* de vulnerabilidad social.

Ello no autoriza a concluir sin más, que toda herramienta actuarial sea discriminatoria, pero si revela un riesgo estructural donde la precariedad puede terminar traducida en peligrosidad y que la intervención más intensa recaiga, no sobre una mayor responsabilidad jurídicamente argumentada, sino sobre trayectorias vitales ya marcadas por la exclusión y sobreintervención institucional. En un sistema de justicia juvenil, esa deriva resulta especialmente problemática, porque transforma condiciones de vulnerabilidad en fundamento indirecto de respuestas más intrusivas, tensionando la especialidad del régimen y la exigencia de una intervención orientada a la reintegración.

El tercer límite dogmático se encuentra referido al alcance mismo de la idea de evidencia cuando se trata de instrumentos actuariales, en rigor el dato de riesgo no describe un hecho pasado ni prueba directamente una conducta expresa una inferencia probabilística construida a partir de variables seleccionadas y ponderadas conforme con determinados supuestos metodológicos. Por ello, su incorporación al proceso decisional no puede presentarse como si se tratara de una constatación neutra o de una verdad autosuficiente.

La fuerza persuasiva de la categoría técnica puede dotar al resultado de una indebida autoridad epistémica, particularmente, cuando los operadores jurídicos carecen de herramientas para reconstruir el razonamiento que lo produce (Bracho-Fuenmayor, 2025b)2025b; por ello, la objeción dogmática no se ve agotada en la opacidad técnica en sentido estricto, sino que alcanza también el riesgo de atribuir a la predicción una fuerza justificativa que excede ontológicamente sus propios límites.

Pues, la evidencia actuarial puede aportar información útil para organizar la intervención, pero en modo alguno puede ser tratada como un fundamento autónomo de legitimación de una respuesta penal juvenil, debido a que su estructura inferencial no satisface por sí sola las exigencias de fundamentación que impone un sistema basado en especialidad. Desde este ángulo, el límite central puede formularse de forma clara, toda vez que, en la justicia penal adolescente, la evidencia actuarial es dogmáticamente admisible solo en la medida que esta permanezca subordinada a la racionalidad jurídica propia del sistema.

Sobre la base de lo *ut supra* esgrimido, no puede sustituir la reconstrucción singular del caso ni el juicio interdisciplinario sobre el adolescente concreto; tampoco puede operar como razón autosuficiente para intensificar la intervención, reducir los márgenes de decisión o justificar respuestas más gravosas, por ello, debe ser tratada como un insumo técnico revisable y no como una conclusión inmune al contraste, toda vez que su uso solo es legítimo si permanece subordinado a la finalidad socioeducativa y de reintegración propia del régimen penal adolescente.

En otras palabras, la tecnificación del sistema no es ilegítima por sí misma, en razón de que lo dogmáticamente inaceptable es que la gestión del riesgo desplace el horizonte normativo de la justicia juvenil y termine convirtiendo una clasificación probabilística en el criterio rector de la decisión.

Sobre esta base, el verdadero problema jurídico no radica en la mera existencia de herramientas actuariales (Kleeven *et al*., 2022), sino en las condiciones bajo las cuales sus resultados pretenden ingresar con relevancia decisional al sistema penal adolescente. Precisamente porque el SEYTD y los instrumentos asociados forman parte de una arquitectura pública, reglada y con incidencia real en la intensidad de la intervención, la discusión ya no pude detenerse en su descripción institucional.

Pues debe desplazarse hacia una cuestión más exigente, relacionada con la transparencia, controlabilidad y contradicción que la prueba puede mantener como para operar legítimamente en un ámbito donde la especialidad, la individualización y el derecho a la defensa no constituyen exigencias periféricas, sino el núcleo mismo de la racionalidad del sistema.

V. Prueba de fiabilidad y derecho a la defensa

La incorporación del informe técnico en la reforma introducida por la Ley 21.527 (Berríos Díaz y Castro Morales, 2025) vuelve ineludible una pregunta procesal que ya no puede ser resuelta solo en el plano de las políticas públicas, y es que bajo cuáles condiciones la evidencia actuarial puede ingresar de forma legítima a la decisión penal adolescente.

El artículo 37 bis de la Ley 20.084 permite que el Ministerio Público o la defensa soliciten al tribunal en cualquier etapa del procedimiento la emisión de un informe técnico a ser evacuado por el Servicio Nacional de Reinserción Social Juvenil, además, dicho informe debe referirse a los criterios del artículo 24 y solo

puede ser utilizado en actuaciones judiciales relativas a la determinación de la sanción una vez evacuado el veredicto condenatorio.

En tal sentido, la propia ley prevé, que sí ninguna de las partes lo solicita, el tribunal puede requerirlo antes de resolver la imposición de la sanción y puede incluso pedir la comparecencia de quienes intervinieron en su preparación en calidad de peritos o requerir su actualización. En consecuencia, el informe técnico (Estrada Vásquez, 2025) no funge como un documento administrativo relevante, sino como un insumo con aptitud legal para incidir en la fase decisoria de la individualización sancionatoria.

Precisamente por ello, el problema central ya no es solo qué tipo de información contiene el informe, sino cuál control puede ejercer la defensa sobre su fundamento técnico, cuando el informe incorpora resultados provenientes de instrumentos estructurados de evaluación del riesgo, la contradicción no puede agotarse en la mera discusión narrativa del caso ni en un contra examen genérico del profesional que lo suscribe.

Si el resultado técnico expresa una clasificación de riesgo con incidencia potencial en la intensidad de la intervención o en la determinación de la sanción, el derecho de defensa exige algo más exigente, la posibilidad real de reconstruir el itinerario que conduce a ese resultado, de cuestionar sus supuestos y de impugnar su valor para el caso concreto.

Desde esta perspectiva, la prueba de fiabilidad no debe concebirse como una exigencia extraña o desproporcionada al sistema, sino como la expresión procesal concreta del derecho a la defensa frente a una evidencia técnica de base probabilística, entonces, en términos dogmáticos, si el resultado actuarial aspira a adquirir relevancia en la decisión, la contradicción debe poder proyectarse al menos sobre cinco dimensiones fundamentales, la validez predictiva, la calibración de resultados, la tasa de error y posibles asimetrías, su validez externa y la eventual existencia de impactos dispares.

Estas categorías no transforman el proceso en una auditoria estadística integral, pero si fijan el contenido mínimo de una contradicción significativa cuando lo que se presenta ante el tribunal no es un mero relato profesional, sino una conclusión construida sobre la base de instrumentos estandarizados.

En este tenor, la experiencia comparada muestra con claridad por qué este punto importa; en *State V. Loomis*, la Corte Suprema de Wisconsin (2016) examinó si la consideración de la evaluación de riesgo COMPAS en la sentencia vulneraba el debido proceso, en particular por la imposibilidad de cuestionar su validez científica debido a su carácter propietario. En ese sentido, el tribunal concluyó que la consideración del instrumento no violaba *per se* el debido proceso, pero solo si esta se usa correctamente, y con conocimiento de sus limitaciones y cautelas; además, subrayó que el puntaje no podía ser determinante en la severidad de la sentencia ni en la decisión sobre la posibilidad de supervisión en comunidad.

El caso si bien no ofrece una solución trasladable sin matices al contexto chileno, si deja una advertencia metodológicamente valiosa, la admisibilidad práctica de este tipo de herramientas depende de que el sistema jurídico no les

atribuya un valor autónomo o concluyente y de que existan advertencias institucionales capaces de impedir que su apariencia técnica sustituya la fundamentación judicial.

En Chile, la cuestión presenta un matiz propio; a diferencia del COMPAS, el SEYTD no aparece configurado como una tecnología privada y comercial, sino como una arquitectura pública y reglada, sin embargo, esa diferencia no elimina el problema de la defensa, simplemente lo reformula. El desafío ya no se concentra únicamente en el acceso al código fuente, sino en la posibilidad de conocer de manera suficiente las variables relevantes, reglas de puntuación, validación del instrumento (Chesta *et al*., 2022), sus límites de uso y el peso real que el resultado tiene en el informe técnico y en la decisión judicial.

Se colige de la tesis esgrimida, que lo advertido ocurre, porque se trata de una herramienta insertada en una institucionalidad pública, el estándar de transparencia exigible no debería ser menos, sino mayor; de ahí que, si la evidencia actuarial ha de operar legítimamente en la determinación de la sanción o en la configuración del plan de intervención, debe hacerlo bajo condiciones que permitan una contradicción técnicamente significativa, esto es, una defensa capaz no solo de alegar en abstracto contra el resultado, sino de discutir de manera informada su fiabilidad, pertinencia y alcance para el caso concreto.

En definitiva, la fiabilidad de la evidencia actuarial se erige como una exigencia procesal ineludible, toda vez que, los principios de espacialidad e individualización propios de la justicia penal adolescente impiden que una clasificación probabilística reemplace la comprensión singular del caso, del mismo modo, el derecho a la defensa exige que sus resultados puedan ser sometidos a contradicción y escrutinio judicial; por ello, su legitimidad no versa solo en su utilidad técnica, sino en su subordinación a una decisión judicial jurídicamente justificada desde la singularidad del caso, y no desde una lógica de mera administración del riesgo.

VI. Estándar de compatibilidad dogmática

La evidencia actuarial solo puede ingresar legítimamente al sistema de responsabilidad del adolescente si permanece estrictamente subordinada a la especialidad (Berríos Díaz, 2011b; Cillero y Vitar, 2022), a la individualización y al derecho a la defensa; de ello se sigue un estándar mínimo de compatibilidad compuesto por cuatro exigencias, trasparencia suficiente para el contradictorio, de modo que la defensa pueda conocer el instrumento empleado, sus variables relevantes, su lógica de clasificación y sus límites; carácter no autosuficiente del resultado, de manera que el puntaje de riesgo no pueda operar como fundamento autónomo de la respuesta judicial; control técnico efectivo que asegure la posibilidad de discutir su fiabilidad, pertenencia y alcance mediante contra examen; así como también una revisión periódica de desempeño y sesgo (Flores, 2024), orientada a detectar errores e impactos indebidos en la población sometida a protección reforzada.

En consecuencia, el problema no radica en la existencia de instrumentos estructurados de evaluación del riesgo, sino en pretender conferirles autonomía decisoria en un ámbito cuya legitimidad exige una respuesta fundada en la singularidad del caso. En el ámbito de la justicia penal adolescente, su uso solo resulta admisible si permanece subordinado a una decisión motivada, individualizada y sometida al contradictorio, de lo contrario, la evidencia actuarial deja de cumplir una función auxiliar y pasa a sustituir el juicio individualizador que impone el ordenamiento jurídico.

VII. Conclusiones

La incorporación de evidencia actuarial en la justicia penal adolescente no puede evaluarse en clave de mera eficiencia técnica, toda vez que su relevancia jurídica depende del lugar que esos resultados ocupen dentro de un sistema cuya legitimidad descansa en la especialidad, la individualización y la protección reforzada de los derechos del adolescente.

En este ámbito, la clasificación del riesgo puede aportar información útil, pero no puede sustituir la justificación singular de la respuesta penal sin desfigurar la racionalidad propia de la justicia juvenil. Se concluye que la evidencia actuarial asociada con la SEYTD no es *per se* incompatible con el ordenamiento jurídico chileno, toda vez que, su límite aparece cuando el resultado técnico deja de operar como insumo revisable y pasa a adquirir valor autónomo en la definición del grado de intervención de la sanción; allí la lógica de predicción invade un espacio que el sistema reserva a la fundamentación jurídica, al juicio interdisciplinario y al contradictorio.

La objeción central, por tanto, no se encuentra dirigida contra la existencia de instrumentos estructurados de evaluación, sino contra su transformación en criterios decisionales autosuficientes. Toda vez que, en el ámbito de la Responsabilidad Penal del Adolescente, ninguna categoría de riesgo puede reemplazar la exigencia de una decisión fundada y que atienda a la singularidad del caso, pues cuando ello ocurre, la técnica deja de asistir al Derecho y comienza a desplazarlo.

Se comprobó que la compatibilidad de la evidencia actuarial con la justicia penal adolescente depende de una condición que vincula su estricta subordinación a la especialidad del sistema, al control contradictorio y a la individualización de la respuesta; pues fuera de ese marco, la promesa de objetividad que acompaña a estos instrumentos no fortalecerá la decisión judicial, sino que la privará de la justificación singular exigida por un régimen orientado por el interés superior del adolescente.

Finalmente se destaca que, el aporte principal de este trabajo radica en afirmar la necesidad de un estándar de compatibilidad dogmática que permita admitir esta evidencia sólo como un insumo técnico revisable, sometido a exigencias reforzadas de transparencia, fiabilidad, contradicción y exclusión de todo valor decisional autónomo, a fin de evitar que la gestión del riesgo desplace la racionalidad jurídica propia de una justicia penal adolescente especializada.

Referencias bibliográficas

Alarcón, P., Pérez-Luco, R., Chesta, S., Wenger, L., Concha-Salgado, A. y García-Cueto, E. (2022). Examining the factor structure of a risk assessment inventory in young offenders: FER-R, Risk and Resource Assessment Form. *International Journal of Environmental Research and Public Health*, *19*(2), 756. https://doi.org/10.3390/ijerph19020756

Argudo, A., Maneiro, L. y Gómez-Fraguela, X. A. (2021). Protocolo de valoración del riesgo en un adolescente en libertad vigilada. *REC. Revista Electrónica de Criminología*, *4*, 1-5. https://doi.org/10.30827/rec.4.33224

Avello Saez, D. M., Zambrano Constanzo, A. X. y Morales, A. R. (2018). Teenagers' personal liability in Chile: Proposals for conducting psychosocial intervention in Juvenile Sections. *Revista Criminalidad*, *60*(3), 205-219. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1794-31082018000300205

Basualto, H. H. (2007). El nuevo derecho penal de adolescentes y la necesaria revisión de su "teoría del delito". *Revista de Derecho (Valdivia)*, *20*(2), 195-217. https://doi.org/10.4067/S0718-09502007000200009

Berríos Díaz, G. (2011). La ley de responsabilidad penal del adolescente como sistema de justicia: Análisis y propuestas. *Política Criminal*, *6*(11), 163-191. https://doi.org/10.4067/S0718-33992011000100006

Berríos Díaz, G. y Castro Morales, Á. (2025). La sustitución de la sanción penal juvenil en la reforma de la Ley Nº 21.527. Propuesta de interpretación en torno al tiempo mínimo de cumplimiento y avances en el plan de intervención. *Política Criminal*, *20*(39), 1-33. https://doi.org/10.4067/s0718-33992025000100001

Berríos, V. S., Villarroel, D. S. y Miranda, L. V. (2025). El uso del interés superior del niño por la Corte Suprema en casos de distribución de medicamentos de alto costo. *Pro Jure Revista de Derecho*, *64*, 1-20. https://doi.org/10.4151/SO2810-76592025064-1527

Bracho-Fuenmayor, P. L. (2025a). Crisis de inteligibilidad probatoria y debido proceso en el Derecho penal: Límites epistémicos de la IA en la imposición de sanciones. *Revista Dike, Irene y Eunomia*, *1*(2), 105-124. https://doi.org/10.36151/RDIE.2025.1.2.06

Bracho-Fuenmayor, P. L. (2025b). Más allá de la exégesis: Reconstrucción dogmática del Derecho penal especial. *Chornancap Revista Jurídica*, *3*(2), 241-253. https://doi.org/10.61542/rjch.165

Bracho-Fuenmayor, P. L. y Pérez Cobo, G. (2025, 4 de abril). El sistema penal juvenil venezolano: Una interpretación desde el pensamiento de Bustos Ramírez. *Revista Argentina de Derecho Público. Edición Especial. Legado y Vanguardia: Explorando el pensamiento penal de Bustos Ramírez en el contexto contemporáneo de Latinoamérica*. https://ijeditores.com/pop.php?option=articulo&Hash=3b684417922ed0c5c887bb52880b3c1e

Carrasco Jiménez, E. (2019). An example of misleading use of the "special" concept by jurisprudence: The definition of "special" from the border of

adolescent criminal responsibility. *Ius et Praxis*, *25*(3), 145-166. https://doi.org/10.4067/S0718-00122019000300145

Chesta, S., Pérez-Luco, R., Alarcón, P., Wenger, L., Concha-Salgado, A. y García-Cueto, E. (2022). Empirical determination of transitory and persistent delinquency in Chilean youth: Validation of the Criminal Engagement Severity Scale "EGED". *International Journal of Environmental Research and Public Health*, *19*(3), 1396. https://doi.org/10.3390/ijerph19031396

Chile. (2005). *Ley 20.084. Establece un sistema de responsabilidad de los adolescentes por infracciones a la ley penal*. https://www.bcn.cl/leychile/navegar?idNorma=244803

Chile. (2022). *Ley 21.430. Sobre garantías y protección integral de los derechos de la niñez y adolescencia*. https://www.bcn.cl/leychile/navegar?idNorma=1173643

Chile. (2023). *Ley 21.527. Crea el Servicio Nacional de Reinserción Social Juvenil e introduce modificaciones a la Ley nº 20.084, sobre responsabilidad penal de adolescentes, y a otras normas que indica*. https://www.bcn.cl/leychile/navegar?idNorma=1187684

Cillero, M. y Vitar, J. (2022). *Responsabilidad penal adolescente*. Academia Judicial de Chile. https://academiajudicial.cl/wp-content/uploads/2024/02/MD60-Resposabilidad-Penal-Adolescente.pdf

Comité de los Derechos del Niño. (2019). *Observación general núm. 24 relativa a los derechos del niño en el sistema de justicia juvenil*. https://www.defensorianinez.cl/wp-content/uploads/2019/12/G1927560.pdf

Corte Suprema de Wisconsin. (2016, 13 de julio). *State v. Loomis*, 2016 WI 68. https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690

Estrada Vásquez, F. (2025). El informe técnico en la reforma a la Ley N° 20.084 de Responsabilidad Penal Adolescente. *Pro Jure Revista de Derecho*, *64*, 1-20. https://doi.org/10.4151/SO2810-76592025064-1486

Flores, Á. C. (2024). Los sesgos algorítmicos en la toma de decisiones automatizadas: Retos y oportunidades para el sistema jurídico peruano. *Informática y Derecho. Revista Iberoamericana de Derecho Informático (2.ª época)*, *2*(15), 159-170. https://revistas.fcu.edu.uy/index.php/informaticayderecho/article/view/5089

González Laurino, C., Guemureman, S. y Leopold Costábile, S. (2024). *Miradas cruzadas. La infracción adolescente y el sistema penal juvenil en Argentina, Brasil, Chile y Uruguay*. FCU. https://libros.fcu.edu.uy/index.php/fcu/es/catalog/book/111

Graepp, F. G. y Ramírez, G. E. (2003). Bases y límites para la responsabilidad penal de los adolescentes. *Revista de Derecho (Valdivia)*, *14*, 99-124. http://revistas.uach.cl/index.php/revider/article/view/2722

Ibarra, J. D. G. y Barragán, L. A. R. (2007). La administración de justicia de menores en México. La reforma del artículo 18 de la Constitución Política de

los Estados Unidos Mexicanos. *Boletín Mexicano de Derecho Comparado*, *50*(148), 207-237. https://doi.org/10.22201/iij.24484873e.2007.118.3907

Kim, K., Duwe, G., Tiry, E., Oglesby-Neal, A., Hu, C., Shields, R., Letourneau, E. y Caldwell, M. (2019). Development and validation of an actuarial risk assessment tool for juveniles with a history of sexual offending. *CrimRxiv*. https://doi.org/10.21428/cb6ab371.66e189b4

Kleeven, A. T. H., De Vries Robbé, M., Mulder, E. A. y Popma, A. (2022). Risk assessment in juvenile and young adult offenders: Predictive validity of the SAVRY and SAPROF-YV. *Assessment*, *29*(2), 181-197. https://doi.org/10.1177/1073191120959740

Malacarne, E. K. y De Azevedo, R. G. (2022). A justiça (penal) juvenil entre a teoria e a prática: Um estudo comparado das práticas judiciais fluminense e gaúcha. *Dilemas - Revista de Estudos de Conflito e Controle Social*, *15*(1), 153-178. https://doi.org/10.4322/dilemas.v15n1.38772

Mettifogo, D., Arévalo, C., Gómez, F., Montedónico, S. y Silva, L. (2015). Factores transicionales y narrativas de cambio en jóvenes infractores de ley: Análisis de las narrativas de jóvenes condenados por la Ley de Responsabilidad Penal Adolescente. *Psicoperspectivas*, *14*(1), 77-88. https://doi.org/10.5027/psicoperspectivas-Vol14-Issue1-fulltext-502

ONU. (1990). *Convención sobre los Derechos del Niño*. https://www.ohchr.org/es/instruments-mechanisms/instruments/convention-rights-child

Ramírez, S. G. (2007). Villanueva, Ruth, Los menores infractores en México. *Boletín Mexicano de Derecho Comparado*, *50*(119), 241-262. https://doi.org/10.22201/iij.24484873e.2007.119.3927

Ramírez, S. G. (2017). Tres ordenamientos del "nuevo sistema penal". Mecanismos alternativos, ejecución de penas y justicia para adolescentes. *Boletín Mexicano de Derecho Comparado*, *50*(149), 1023-1043. https://doi.org/10.22201/iij.24484873e.2017.149.11365

Salas, J. C. (2012). Los adolescentes ante el Derecho penal en Chile. Estándares de juzgamiento diferenciado en materia penal sustantiva. *Revista de Derecho (Valdivia)*, *25*(1), 149-173. https://doi.org/10.4067/S0718-09502012000100007

Sánchez, N. B. B. y Lescano-Galeas, N. (2021). Resignificar la justicia penal. Un análisis entre la práctica de Ecuador y México. *Boletín Mexicano de Derecho Comparado*, *54*(162), 251-290. https://doi.org/10.22201/iij.24484873e.2021.162.17072

Torres-Vásquez, H. y Tirado-Acero, M. (2023). Las sanciones en el Sistema de Responsabilidad Penal Adolescente en Colombia. *Revista Científica General José María Córdova*, *21*(41), 131-148. https://doi.org/10.21830/19006586.1001

Velásquez, C. M., Carrillo, O. G. y Pacheco, B. G. (2023). Modelos, sanciones y desarrollo de la finalidad educativa en el Sistema de Responsabilidad Penal para Adolescentes. Un análisis en el adolescente infractor del delito de hurto en la ciudad de Barranquilla. *Revista Criminalidad*, *65*(1), 27-40. https://doi.org/10.47741/17943108.399

Vera Vega, J. y Mayer Lux, L. (2024). La nueva regulación del quebrantamiento de condena de adolescentes: Naturaleza, sentido y vigencia. *Revista Chilena de Derecho*, *51*(3), 133-162. https://doi.org/10.7764/R.513.5

Zambrano, V., Fernández-Pacheco, G. y Salazar-Muñoz, M. (2023). The transition of Chilean adolescents from the child welfare system to the adolescent justice system: A continuation or an accumulation of adverse factors? *Frontiers in Psychology*, *14*, 1194294. https://doi.org/10.3389/fpsyg.2023.1194294

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 224-251

EXPLICABILIDAD Y TRAZABILIDAD EN LA INTELIGENCIA ARTIFICIAL GENERATIVA EN EL ECOSISTEMA JURÍDICO: HACIA UN MODELO DE GOBERNANZA ORIENTADO A DERECHOS

EXPLAINABILITY AND TRACEABILITY IN GENERATIVE AI FOR THE LEGAL ECOSYSTEM: A RIGHTS-ORIENTED GOVERNANCE FRAMEWORK

Iván Miguel García-López y Laura Trujillo-Liñán
Universidad Panamericana

Resumen

La incorporación de inteligencia artificial generativa en el ecosistema jurídico ha transformado la producción, revisión y validación del trabajo legal, generando tensiones entre innovación tecnológica, protección de derechos, diligencia profesional y responsabilidad institucional. Este artículo analiza si la explicabilidad constituye un estándar suficiente para gobernar el uso de estas herramientas en contextos jurídicos. Se sostiene que, aunque necesaria, la explicabilidad resulta insuficiente como mecanismo autónomo, ya que no permite reconstruir el proceso de generación, verificar la corrección jurídica del contenido ni atribuir responsabilidad por errores o usos indebidos. Metodológicamente, el estudio adopta un enfoque jurídico-dogmático y propositivo. Para ello, examina la transformación del ecosistema jurídico, sistematiza el marco normativo aplicable, analiza críticamente los límites de la explicabilidad y desarrolla una matriz jurídica de uso legítimo de IA generativa. Como principal aportación, se propone una matriz operativa estructurada en seis dimensiones: delimitación del caso de uso, gobernanza de datos, trazabilidad documental, supervisión humana significativa, gestión del riesgo y responsabilidad documentable. A diferencia de marcos éticos generales, la matriz traduce principios normativos en indicadores, evidencias, responsables y ejemplos de aplicación práctica. Asimismo, el artículo incorpora objeciones doctrinales y límites prácticos relativos al coste de implementación, el secreto profesional, la confidencialidad y el riesgo de formalismo documental. Se concluye que la IA generativa solo puede considerarse jurídicamente legítima cuando su uso es trazable, supervisado, proporcional al riesgo y jurídicamente imputable.

Palabras clave

IA generativa, servicios jurídicos, trazabilidad documental algorítmica, gobernanza jurídica de IA, supervisión humana significativa, responsabilidad profesional documentable

Abstract

The incorporation of generative artificial intelligence into the legal ecosystem has transformed the production, review, and validation of legal work, generating tensions between technological innovation, rights protection, professional diligence, and institutional accountability. This article examines whether explainability constitutes a sufficient standard for governing the use of these tools in legal contexts. It argues that, although necessary, explainability is insufficient as an autonomous mechanism, since it does not allow for reconstructing the generation process, verifying the legal correctness of the content, or attributing responsibility for errors or misuse. Methodologically, the study adopts a legal-dogmatic and propositional approach. It examines the transformation of the legal ecosystem, systematizes the applicable normative framework, critically analyzes the limits of explainability, and develops a legal matrix for the legitimate use of generative AI. As its main contribution, the article proposes an operational matrix structured around six dimensions: use case delimitation, data

governance, document traceability, meaningful human oversight, risk management, and documentable accountability. Unlike general ethical frameworks, the matrix translates normative principles into indicators, evidence, responsible actors, and practical examples of application. The article also addresses doctrinal objections and practical limits related to implementation costs, professional secrecy, confidentiality, and the risk of documentary formalism. It concludes that generative AI can only be considered legally legitimate when its use is traceable, supervised, proportionate to risk, and legally accountable.

Keywords

Generative AI, legal services, algorithmic document traceability, legal AI governance, meaningful human oversight, documentable professional accountability

I. Introducción

La expansión de la inteligencia artificial generativa ha transformado de manera profunda el ecosistema jurídico contemporáneo. Despachos, tribunales y organizaciones jurídicas utilizan estas herramientas para redactar, analizar y estructurar información jurídica, lo que modifica no solo la eficiencia del trabajo, sino las condiciones bajo las cuales se produce y valida el razonamiento jurídico (Felin y Holweg, 2024). En este contexto, el problema central ya no es únicamente su utilidad, sino en qué condiciones su uso puede considerarse jurídicamente legítimo y compatible con la protección de derechos (Grimmelikhuijsen y Meijer, 2022).

A diferencia de otros sectores, el ámbito jurídico opera con categorías de alta densidad institucional como diligencia profesional, confidencialidad, motivación y responsabilidad, por lo que los riesgos asociados a la IA generativa adquieren una intensidad particular (Ahmad *et al*., 2023). La opacidad de los modelos, la posibilidad de errores plausibles y la dificultad para reconstruir el proceso de generación plantean desafíos específicos de explicabilidad y trazabilidad, que afectan la supervisión, la verificación y la atribución de responsabilidad (Herrera y Calderón, 2025).

El artículo sostiene que la explicabilidad, aunque necesaria, resulta insuficiente como estándar de control (Altukhi y Pradhan, 2025). En su lugar, propone un modelo de gobernanza orientado a derechos, basado en trazabilidad, supervisión humana significativa, minimización de datos y responsabilidad institucional documentable.

Metodológicamente, se adopta un enfoque jurídico-dogmático y propositivo. El trabajo analiza la transformación del ecosistema jurídico, examina críticamente la explicabilidad, sistematiza el marco normativo aplicable y desarrolla una matriz de gobernanza jurídicamente operativa para el uso legítimo de IA generativa (Bauer *et al*., 2023; Floridi, 2016). Esta matriz se construye a partir de la articulación entre protección de datos, transparencia, supervisión humana, trazabilidad, gestión del riesgo y responsabilidad documentable, incorporando

también objeciones doctrinales y límites prácticos relativos al coste de implementación, el secreto profesional, la confidencialidad y el riesgo de formalismo documental.

II. IA generativa y transformación del ecosistema jurídico

La inteligencia artificial generativa ha reconfigurado de manera sustantiva el ecosistema jurídico al intervenir directamente en la producción, organización y validación del trabajo legal. A diferencia de sistemas tradicionales de automatización, los modelos generativos participan en tareas que inciden en el núcleo de la práctica jurídica, como la redacción de documentos, el análisis de casos, la estructuración argumentativa y la síntesis normativa (Adiguzel *et al*., 2023). Esto desplaza parcialmente el proceso de elaboración jurídica hacia sistemas capaces de generar resultados plausibles y persuasivos, lo que modifica las condiciones de confianza, validación y responsabilidad en el ejercicio profesional (Afroogh *et al*., 2024).

El concepto de ecosistema jurídico permite comprender que este impacto no se limita al trabajo individual, sino que atraviesa una red institucional compuesta por despachos, tribunales, organizaciones públicas, universidades y usuarios de servicios jurídicos (Aleem, 2026). En este entorno, la IA generativa no solo altera tareas, sino también estándares de diligencia, formas de supervisión y criterios de control, al introducir dinámicas de producción híbrida entre humanos y sistemas algorítmicos (Awasthi *et al*., 2025).

1. De herramienta de apoyo a infraestructura cognitiva

Un cambio central es la transición de la IA generativa de herramienta auxiliar a infraestructura cognitiva del trabajo jurídico. Mientras que tecnologías previas facilitaban la búsqueda o gestión de información, los modelos generativos reorganizan, interpretan y presentan contenido en forma argumentativa, desplazando el valor profesional hacia la validación, corrección y control de resultados generados algorítmicamente (Ananny y Crawford, 2018). En consecuencia, el profesional deja de ser únicamente productor de contenido y asume un rol de supervisor de procesos híbridos (Cornejo-Plaza y Cippitani, 2023).

Este desplazamiento tiene implicaciones normativas relevantes. El lenguaje generado por estos sistemas puede ser jurídicamente significativo y persuasivo incluso cuando contiene errores, lo que exige reforzar los mecanismos de verificación y control (Cotton *et al*., 2023). La cuestión ya no es solo la eficiencia, sino la fiabilidad del contenido en contextos donde el error puede traducirse en consecuencias jurídicas concretas (Bathaee, 2018).

2. Reconfiguración de deberes profesionales

La integración de IA generativa modifica los deberes profesionales al exigir una combinación de competencia jurídica y competencia tecnológica (Bazanova *et al*., 2023). El uso de estos sistemas implica no solo comprender sus capacidades,

sino también sus límites, riesgos y condiciones de uso adecuado. En particular, se intensifican los deberes de verificación, confidencialidad y supervisión, ya que la confianza acrítica en resultados generados puede comprometer la calidad del servicio jurídico (Abhishek *et al*., 2025).

En este contexto, la responsabilidad profesional no se reduce, sino que se amplía. El uso de IA generativa exige validar el contenido producido, proteger los datos utilizados y garantizar que el resultado final cumple con estándares jurídicos y éticos (Alvarado, 2023). Esto implica que los errores derivados del uso de estos sistemas no pueden considerarse externos al profesional, sino parte de su esfera de responsabilidad (Birkstedt *et al*., 2023).

3. Riesgos estructurales

La IA generativa introduce riesgos específicos que adquieren especial relevancia en el ámbito jurídico. Entre ellos destacan la producción de contenido incorrecto pero plausible, la opacidad del proceso de generación, la posible exposición de datos sensibles y la reducción del escrutinio crítico por parte del usuario (Ferrara, 2023). Estos riesgos no actúan de manera aislada, sino que se potencian mutuamente, generando escenarios en los que la calidad y la confiabilidad del trabajo jurídico pueden verse comprometidas (Ferrer *et al*., 2021).

En particular, la combinación de opacidad y plausibilidad aumenta la dificultad de detectar errores, mientras que la falta de control sobre los datos introducidos puede afectar la confidencialidad y la protección de información sensible (Kelly et al., 2023). Asimismo, la dependencia de sistemas generativos puede desplazar el esfuerzo de validación hacia etapas posteriores, incrementando la carga de verificación (Breen *et al*., 2020).

4. De la adopción instrumental a la gobernanza jurídica

Como consecuencia, la IA generativa no puede ser tratada como una herramienta neutra de apoyo, sino como un objeto de gobernanza jurídica (Silva, 2022). Su uso plantea la necesidad de establecer reglas sobre documentación, supervisión, trazabilidad, protección de datos y distribución de responsabilidades (Bano *et al*., 2023). En este sentido, el problema central deja de ser si la tecnología puede utilizarse, para convertirse en bajo qué condiciones su uso resulta jurídicamente legítimo.

Este desplazamiento del análisis desde la utilidad técnica hacia la gobernanza prepara el siguiente paso del argumento: examinar los alcances y límites de la explicabilidad como categoría central en la regulación de la inteligencia artificial generativa en el ecosistema jurídico (Ananny y Crawford, 2018).

III. Explicabilidad en IA generativa: alcance y límites

La explicabilidad se ha consolidado como una de las nociones centrales en el debate contemporáneo sobre inteligencia artificial, particularmente en contextos de alta relevancia institucional (Floridi, 2023). En el ecosistema jurídico,

su importancia es aún mayor, dado que las salidas generadas por modelos de lenguaje pueden integrarse en productos jurídicos con efectos prácticos, como escritos, contratos, análisis o decisiones (Ben-Michael *et al*., 2024). Bajo estas condiciones, exigir explicabilidad parece una respuesta natural frente a la opacidad de los sistemas generativos.

Sin embargo, la explicabilidad en IA generativa presenta limitaciones estructurales. A diferencia de modelos predictivos más acotados, los sistemas generativos producen secuencias textuales a partir de procesos probabilísticos complejos, lo que dificulta traducir su funcionamiento interno en razones útiles para control jurídico, revisión profesional o atribución de responsabilidad (Floridi *et al*., 2020). En este sentido, una explicación técnica del modelo no necesariamente permite comprender por qué se generó una salida específica ni evaluar su corrección jurídica (Bearman y Ajjawi, 2023).

1. La promesa de la explicabilidad

La explicabilidad ofrece una promesa de inteligibilidad frente a sistemas que producen resultados plausibles sin revelar su proceso de generación (Floridi *et al*., 2018). En el ámbito jurídico, esta promesa resulta particularmente atractiva porque podría facilitar la identificación de errores, la verificación de contenido y la evaluación del grado de confianza en una salida (Rambachan, 2024).

No obstante, su valor es limitado cuando se reduce a una función informativa. Una explicación del funcionamiento del sistema no equivale a una validación del contenido generado ni garantiza la posibilidad de revisión efectiva (Atenas, 2024). Por ello, la explicabilidad debe entenderse como un punto de partida, no como un estándar suficiente de control (Zhang *et al*., 2024).

2. Límites estructurales de la explicabilidad

Los modelos generativos presentan al menos tres límites relevantes. En primer lugar, las explicaciones simplificadas suelen ser aproximaciones narrativas que no reflejan fielmente el proceso interno del modelo (Guidotti *et al*., 2019). En segundo lugar, las explicaciones generadas por el propio sistema pueden constituir nuevas salidas probabilísticas, sin garantía de correspondencia con el proceso real (Gangavarapu, 2025). En tercer lugar, la comprensión general del modelo no permite reconstruir una interacción concreta ni verificar el origen de los elementos incorporados en la salida.

Estos límites adquieren especial relevancia en el ámbito jurídico, donde el lenguaje generado puede tener apariencia de completitud y autoridad (Hasibuan *et al*., 2024). En tales casos, la explicabilidad no sustituye la necesidad de validación externa ni garantiza la corrección normativa del contenido (Méndez Carpio *et al*., 2023).

3. Niveles de explicabilidad

Para precisar su alcance, es útil distinguir tres niveles de explicabilidad. El primero es la explicabilidad técnica, orientada a comprender el funcionamiento general del modelo (Clark *et al*., 2020). El segundo es la explicabilidad funcional, que permite identificar cómo intervino el sistema en una tarea específica (Hassoulas *et al*., 2023). El tercero es la explicabilidad jurídica, que traduce esa intervención en términos normativos relevantes, como responsabilidad, deberes aplicables y posibilidades de control (Unión Europea, 2024).

La clave es que estos niveles no se implican automáticamente. Un sistema puede ser comprensible en términos técnicos y funcionales, y aun así resultar insuficiente desde el punto de vista jurídico si no permite validar, reconstruir o atribuir responsabilidad por el contenido generado (Bauer *et al*., 2023).

4. Explicabilidad y gobernanza

La principal limitación de la explicabilidad es que puede ofrecer comprensión sin permitir reconstrucción verificable (Schneegans et al., 2021). Una explicación puede describir el funcionamiento del sistema, pero no documenta necesariamente qué ocurrió en una interacción concreta, qué insumos se utilizaron o qué intervención humana existió (Leung et al., 2023).

Por ello, la explicabilidad debe integrarse en un marco más amplio de gobernanza. Mientras que la explicación responde parcialmente al «por qué» de una salida, la trazabilidad permite responder al «cómo», «con qué» y «bajo qué responsabilidad» se produjo (Díaz-Rodríguez et al., 2023). Esta distinción es decisiva en el ámbito jurídico, donde la legitimidad del uso de IA generativa depende de la posibilidad de someter sus resultados a control, revisión y atribución de responsabilidad (Al-Kfairy et al., 2024).

En consecuencia, la explicabilidad constituye una condición necesaria, pero no suficiente, para gobernar jurídicamente la inteligencia artificial generativa (De Oliveira Fornasier, 2021). Su función principal es habilitar niveles mínimos de inteligibilidad que permitan activar mecanismos más robustos de control, particularmente aquellos vinculados con la trazabilidad y la responsabilidad institucional (Núñez Portilla y Guerrero Zambrano, 2025). Esta conclusión conduce al siguiente eje del análisis: la trazabilidad como elemento central para la reconstrucción, auditoría y control del uso de IA generativa en el ecosistema jurídico.

IV. Trazabilidad y cadena de responsabilidad

La trazabilidad puede definirse como la capacidad de reconstruir de manera verificable el proceso de intervención de la inteligencia artificial generativa en una actividad jurídica concreta. No se trata únicamente de registrar aspectos técnicos del sistema, sino de identificar qué herramienta se utilizó, qué insumos se introdujeron, qué salida se generó, qué intervención humana existió y cómo ese contenido se incorporó a un producto jurídico final (Königs, 2022). En el

ecosistema jurídico, donde los resultados tienen efectos normativos, procesales o patrimoniales, esta reconstrucción es esencial para sostener verificación, auditoría y atribución de responsabilidad (Casey et al., 2019).

1. Trazabilidad como presupuesto de gobernanza

La trazabilidad constituye un requisito básico de gobernanza jurídica (Schneegans et al., 2021). A diferencia de la explicabilidad, que aporta comprensión general, la trazabilidad permite documentar el uso concreto del sistema, facilitando verificación interna, auditoría institucional y delimitación de responsabilidades (Hogan et al., 2022). Sin registros adecuados, el uso de IA generativa pierde control jurídico y dificulta demostrar diligencia profesional o cumplimiento normativo.

Este aspecto resulta particularmente relevante en contextos de presión operativa, donde las herramientas generativas pueden incorporarse rápidamente a flujos de trabajo sin reglas claras de documentación (OpenAI et al., 2023). En tales casos, la organización puede ganar eficiencia aparente, pero pierde capacidad de reconstrucción, lo que compromete su posición frente a errores, conflictos de confidencialidad o cuestionamientos sobre la calidad del servicio jurídico (Veale y Zuiderveen Borgesius, 2021).

2. Elementos mínimos de trazabilidad

Para que la trazabilidad tenga valor jurídico, debe incluir al menos cinco elementos: identificación del sistema utilizado, registro de entradas (*prompts* y documentos), salida generada, intervención humana posterior y destino del producto final (Choudhary et al., 2025). Esta cadena permite distinguir entre generación automática y validación profesional, y constituye la base para revisión y control (Association for Computing Machinery, 2018).

Adicionalmente, resulta indispensable documentar el tratamiento de datos sensibles o confidenciales (Kelly et al., 2023). En el ámbito jurídico, los insumos pueden incluir información protegida por secreto profesional o normativa de protección de datos, por lo que la trazabilidad debe abarcar qué datos se utilizaron, bajo qué criterios de necesidad y con qué medidas de protección (Alshami et al., 2023).

3. Cadena de responsabilidad en entornos híbridos

La trazabilidad se complementa con la delimitación de la cadena de responsabilidad en entornos híbridos humano-IA (Cheong, 2024). En estos contextos, la producción jurídica suele involucrar múltiples actores: proveedores del sistema, organizaciones que lo adoptan, usuarios que lo operan y profesionales que validan el resultado (Marín y Tur, 2023). Sin una estructura clara, esta distribución puede generar responsabilidad difusa.

Desde una perspectiva jurídica, la intervención de IA generativa no elimina la responsabilidad, sino que la redistribuye (Golovianko et al., 2023). En

particular, pueden distinguirse cuatro niveles: responsabilidad de diseño (proveedor), responsabilidad de implementación (organización), responsabilidad de uso (usuario) y responsabilidad de validación final (quien adopta el producto jurídico) (Trikoili et al., 2025). Esta diferenciación permite evitar tanto la dilución de responsabilidades como su atribución simplificada.

4. Trazabilidad, prueba y auditabilidad

La trazabilidad cumple también una función probatoria. En contextos de controversia, auditoría o revisión institucional, la capacidad de reconstruir el uso de IA generativa permite demostrar diligencia, identificar errores y evaluar responsabilidades (Theis et al., 2023). Sin registros, la reconstrucción depende de versiones retrospectivas; con trazabilidad, el uso del sistema se vuelve auditable (Schneegans et al., 2021).

En este sentido, la trazabilidad conecta la innovación tecnológica con categorías jurídicas tradicionales como prueba, control y responsabilidad (Agbese et al., 2023). Su valor no es meramente técnico, sino institucional: permite someter el uso de IA generativa a estándares verificables compatibles con el ejercicio del Derecho.

En consecuencia, la trazabilidad constituye el puente entre explicabilidad y gobernanza jurídica. Mientras que la explicación permite comprender el sistema, la trazabilidad permite controlarlo (Bach et al., 2024). Esta distinción resulta decisiva para establecer condiciones de uso legítimo de IA generativa en el ecosistema jurídico, lo que conduce al desarrollo de un modelo de gobernanza orientado a derechos.

V. Gobernanza orientada a derechos

La integración de inteligencia artificial generativa en el ecosistema jurídico exige una forma de gobernanza que supere enfoques meramente técnicos o voluntaristas (Barocas et al., 2018). Dado que estos sistemas intervienen en la producción de lenguaje jurídicamente relevante, su uso puede afectar directamente la confidencialidad, la calidad del servicio, la responsabilidad profesional y la protección de derechos (Bellamy, 2017). En consecuencia, la gobernanza debe estructurarse a partir de estándares jurídicos orientados a garantizar condiciones de uso legítimo y controlable.

En este contexto, la gobernanza orientada a derechos puede definirse como el conjunto de reglas y procedimientos destinados a asegurar que el uso de IA generativa sea compatible con exigencias de protección de datos, diligencia profesional, supervisión humana, control del riesgo y responsabilidad institucional (Head et al., 2023). Su foco no recae únicamente en el sistema, sino en el entorno de uso, considerando que los riesgos varían según el contexto institucional y la función jurídica desempeñada.

1. Marco normativo aplicable

La gobernanza jurídica de la inteligencia artificial generativa no puede construirse únicamente desde principios éticos generales, sino que requiere identificar las obligaciones positivas que ya estructuran el uso institucional de sistemas algorítmicos. En el plano europeo, el Reglamento General de Protección de Datos establece principios de licitud, lealtad, transparencia, limitación de finalidad, minimización, exactitud, seguridad y responsabilidad proactiva, lo que vincula directamente el uso de IA generativa con deberes de justificación, documentación y control del tratamiento de datos personales (Unión Europea, 2016). A su vez, el RGPD reconoce restricciones específicas frente a decisiones basadas únicamente en tratamiento automatizado con efectos jurídicos o significativamente similares, e incorpora la intervención humana como salvaguarda mínima en determinados supuestos (Unión Europea, 2016). Desde una lógica organizacional, el deber de implementar medidas técnicas y organizativas apropiadas, así como registros de actividades de tratamiento, permite conectar protección de datos, trazabilidad y responsabilidad institucional (Unión Europea, 2016). Por ello, la matriz propuesta no introduce un estándar externo al Derecho positivo, sino que operacionaliza obligaciones ya reconocibles en materia de protección de datos, supervisión y rendición de cuentas.

El Reglamento Europeo de Inteligencia Artificial refuerza esta lectura al adoptar un enfoque basado en riesgos orientado a promover una IA centrada en la persona, confiable y compatible con los derechos fundamentales, la democracia y el Estado de Derecho (Unión Europea, 2024). Este marco distingue sistemas de alto riesgo, exige documentación técnica y contempla obligaciones de transparencia para determinados sistemas de IA, incluidos supuestos de interacción con personas o generación de contenido sintético (Unión Europea, 2024). Aunque no todo uso jurídico de IA generativa será automáticamente de alto riesgo, la racionalidad regulatoria resulta trasladable al ecosistema jurídico: cuanto mayor sea el impacto sobre derechos, confidencialidad, estrategia procesal o consecuencias patrimoniales, mayor debe ser la intensidad de documentación, supervisión y control. En este sentido, la matriz propuesta traduce el enfoque basado en riesgos a prácticas jurídicas concretas, especialmente cuando el sistema se utiliza para redactar, sintetizar, evaluar o apoyar productos jurídicos susceptibles de afectar intereses de terceros.

Este marco se complementa con instrumentos internacionales de derechos humanos y ética pública de la IA. El Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho exige que las actividades dentro del ciclo de vida de los sistemas de IA sean compatibles con derechos humanos, democracia y Estado de Derecho (Consejo de Europa, 2024). La Recomendación de la Unesco sobre la Ética de la Inteligencia Artificial, por su parte, sitúa la dignidad humana, los derechos humanos, la transparencia, la equidad y la supervisión humana como condiciones estructurales para orientar políticas públicas y marcos institucionales de IA (Unesco, 2021). Leídos conjuntamente, estos instrumentos permiten sostener que la trazabilidad, la supervisión humana significativa y la responsabilidad documentable no son simples preferencias de buena práctica, sino mecanismos necesarios

para traducir principios de derechos humanos en rutinas institucionales verificables. En consecuencia, la matriz jurídica propuesta debe entenderse como una herramienta intermedia entre los principios normativos generales y las decisiones concretas de implementación en despachos, tribunales, áreas jurídicas institucionales y entidades públicas. La Tabla 1 sistematiza esta conexión entre fuentes normativas e institucionales, exigencias jurídicas relevantes y traducción operativa dentro de la matriz propuesta.

Tabla 1. Marco normativo aplicable al uso de IA generativa en el ecosistema jurídico

Fuente normativa o institucional	Exigencia relevante	Traducción operativa en la matriz propuesta
Reglamento General de Protección de Datos, Reglamento (UE) 2016/679	Licitud, transparencia, limitación de finalidad, minimización, seguridad, responsabilidad proactiva, intervención humana frente a determinadas decisiones automatizadas y registro de actividades de tratamiento	Gobernanza de datos, minimización, trazabilidad documental y responsabilidad documentable
Reglamento Europeo de Inteligencia Artificial, Reglamento (UE) 2024/1689	Enfoque basado en riesgos, documentación técnica, transparencia, supervisión humana y protección de derechos fundamentales	Clasificación del caso de uso, gestión del riesgo, supervisión humana significativa y evidencia de control
Convenio Marco del Consejo de Europa sobre Inteligencia Artificial, Derechos Humanos, Democracia y Estado de Derecho	Compatibilidad de los sistemas de IA con derechos humanos, democracia y Estado de Derecho durante el ciclo de vida del sistema	Trazabilidad, auditabilidad, revisión institucional y mecanismos de impugnación o corrección
Recomendación de la UNESCO sobre la Ética de la Inteligencia Artificial	Dignidad humana, derechos humanos, equidad, transparencia, supervisión humana y gobernanza ética de la IA	Gobernanza orientada a derechos, control humano sustantivo y criterios de responsabilidad institucional
Deberes profesionales del ejercicio jurídico	Competencia, diligencia, confidencialidad, verificación de fuentes, secreto profesional y responsabilidad por el producto jurídico final	Supervisión jurídica del resultado, protección de información confidencial y responsabilidad profesional documentable

2. Protección de datos y confidencialidad

La protección de datos constituye un eje estructural de la gobernanza en el ámbito jurídico. El uso de IA generativa puede implicar el tratamiento de información sensible, documentos confidenciales o datos protegidos, lo que exige criterios estrictos de minimización, anonimización y control de acceso (Bano et al., 2023). No toda información jurídicamente disponible debe ser introducida en el

sistema, especialmente cuando su tratamiento no puede garantizar condiciones adecuadas de seguridad o confidencialidad (Ghimire y Edwards, 2024).

En este sentido, la gobernanza debe incorporar reglas claras sobre qué datos pueden utilizarse, bajo qué condiciones y con qué medidas de protección (Choudhary et al., 2025). Esta exigencia no solo responde a obligaciones normativas en materia de datos personales, sino también a deberes éticos fundamentales del ejercicio jurídico.

3. Supervisión humana significativa

El uso de IA generativa exige una supervisión humana que vaya más allá de la revisión formal. La validación debe implicar la verificación sustantiva del contenido, incluyendo su corrección normativa, coherencia argumentativa y adecuación al caso concreto (Bazanova et al., 2023). Dado que estos sistemas pueden producir resultados plausibles pero incorrectos, la supervisión constituye un elemento central para evitar errores jurídicamente relevantes y para alinear el uso institucional de IA generativa con exigencias de control humano, gestión del riesgo y responsabilidad documentable (Unión Europea, 2024; OECD, 2023).

En este contexto, la supervisión debe entenderse como un deber reforzado de validación. El profesional no solo debe revisar el contenido generado, sino también asumir responsabilidad por su uso, lo que implica la capacidad efectiva de corregir, modificar o descartar la salida del sistema (Allaham y Diakopoulos, 2024).

4. Gestión del riesgo y distribución de responsabilidades

La gobernanza orientada a derechos requiere una gestión diferenciada del riesgo, basada en la identificación de escenarios de uso y su potencial impacto jurídico (OECD, 2023). No todas las aplicaciones de IA generativa presentan el mismo nivel de riesgo, por lo que resulta necesario distinguir entre usos admisibles, restringidos y no admisibles (Feldvari et al., 2022).

A esta clasificación debe añadirse una distribución clara de responsabilidades entre los distintos actores involucrados: proveedores, organizaciones y usuarios (Kreps y Kriner, 2023). Esta arquitectura evita la dilución de responsabilidades y permite establecer criterios de imputación acordes con la naturaleza híbrida de los procesos jurídico-tecnológicos.

5. Derechos fundamentales y acceso a la justicia

La gobernanza no puede limitarse a consideraciones internas de eficiencia o control organizacional (Casey et al., 2019). La incorporación de IA generativa en el ecosistema jurídico también plantea implicaciones para derechos fundamentales, como la igualdad, el acceso a la justicia y la tutela efectiva (Mantelero, 2024).

Estas herramientas pueden ampliar el acceso a información jurídica, pero también generar riesgos de sesgo, desinformación, dependencia tecnológica o

afectación diferenciada sobre usuarios con menor capacidad de validación. Por ello, la gobernanza debe incorporar criterios que permitan evaluar su impacto sobre distintos grupos de usuarios, garantizar mecanismos de corrección y revisión, y asegurar que el uso de IA generativa no debilite la igualdad, la tutela efectiva ni la confianza institucional en el sistema jurídico (Consejo de Europa, 2024; Mantelero, 2024; Unesco, 2021).

Estas herramientas pueden ampliar el acceso a información jurídica, pero también generar riesgos de sesgo, desinformación, dependencia tecnológica o afectación diferenciada sobre usuarios con menor capacidad de validación. Por ello, la gobernanza debe incorporar criterios que permitan evaluar su impacto sobre distintos grupos de usuarios, garantizar mecanismos de corrección y revisión, y asegurar que el uso de IA generativa no debilite la igualdad, la tutela efectiva ni la confianza institucional en el sistema jurídico (Consejo de Europa, 2024; Mantelero, 2024; Unesco, 2021).

6. Hacia una gobernanza operativa

En conjunto, la gobernanza orientada a derechos propuesta en este trabajo se articula en un conjunto integrado de exigencias: protección de datos, confidencialidad, supervisión humana significativa, gestión del riesgo, trazabilidad documental, responsabilidad institucional y control proporcional del uso. Su valor radica en traducir principios generales en criterios operativos que permitan controlar el uso de IA generativa en contextos jurídicos (Floridi & Cowls, 2019).

Este enfoque permite integrar las dimensiones de explicabilidad, trazabilidad, protección de datos y responsabilidad en una arquitectura coherente de gobernanza jurídicamente situada (García López et al., 2024). A diferencia de una aproximación puramente ética o conceptual, la lectura sistemática del RGPD, del Reglamento Europeo de Inteligencia Artificial, del Convenio Marco del Consejo de Europa y de la Recomendación de la Unesco permite identificar un núcleo común de exigencias: minimización de datos, transparencia, supervisión humana, gestión del riesgo, documentación y responsabilidad. Sobre esta base, es posible formular una propuesta más concreta: una matriz jurídica de uso legítimo que transforme estos estándares normativos en parámetros aplicables a la práctica jurídica.

VI. Propuesta de matriz jurídica de uso legítimo de ia generativa en el ecosistema jurídico

A partir del análisis previo, el uso jurídicamente legítimo de la inteligencia artificial generativa no puede depender de una sola condición, como la explicabilidad, sino de una arquitectura integrada de gobernanza. Esta debe ser lo suficientemente precisa para orientar la práctica jurídica y lo suficientemente flexible para adaptarse a distintos contextos institucionales (Bandi et al., 2023).

En este sentido, se propone una matriz de uso legítimo basada en seis dimensiones acumulativas: delimitación del caso de uso, gobernanza de datos, trazabilidad, supervisión humana significativa, gestión de riesgo y responsabilidad

documentable (Tarun et al., 2025). Su objetivo es establecer un estándar mínimo que permita evaluar cuándo el uso de IA generativa resulta jurídicamente aceptable.

La matriz propuesta se diferencia de los marcos generales de ética, transparencia o cumplimiento algorítmico en tres aspectos. Primero, no se limita a formular principios abstractos de confianza, explicabilidad o responsabilidad, sino que traduce esos principios en condiciones verificables de uso dentro de flujos jurídicos concretos. Segundo, no clasifica el riesgo únicamente desde la perspectiva técnica del sistema, sino desde la función jurídica que cumple la herramienta en cada caso de uso, distinguiendo entre actividades exploratorias, tareas auxiliares y usos de alto impacto jurídico. Tercero, no concibe la explicabilidad como un fin autosuficiente, sino como una condición instrumental que debe articularse con trazabilidad documental, supervisión humana significativa y responsabilidad profesional documentable. De este modo, la matriz no pretende sustituir marcos regulatorios existentes, sino ofrecer una herramienta intermedia de implementación para despachos, tribunales, áreas jurídicas institucionales, clínicas legales y entidades públicas que empleen IA generativa en actividades de redacción, síntesis, análisis o apoyo jurídico.

1. Delimitación del caso de uso

La primera dimensión consiste en definir el tipo de uso permitido. No todas las tareas jurídicas presentan el mismo nivel de riesgo, por lo que resulta necesario distinguir entre usos exploratorios, usos de apoyo profesional y usos de alto impacto jurídico (Floridi, 2016).

Los usos exploratorios incluyen actividades de ideación o reformulación sin efectos directos. Los usos de apoyo comprenden tareas como borradores o síntesis, siempre sujetos a revisión intensiva. Los usos de alto impacto incluyen la incorporación de contenido en productos jurídicos finales, lo que exige mayores niveles de control (Grünewald y Pallas, 2021). Esta clasificación permite aplicar criterios de proporcionalidad en la gobernanza.

2. Gobernanza de datos

La segunda dimensión se refiere al tratamiento de datos. El uso de IA generativa en contextos jurídicos puede implicar información sensible, por lo que debe regirse por una lógica de minimización reforzada (Zhang et al., 2024). Esto implica utilizar únicamente los datos estrictamente necesarios, priorizar anonimización y restringir el uso de información protegida.

Desde un punto de vista operativo, el uso de la herramienta debe condicionarse a la evaluación previa de la necesidad del dato, la posibilidad de anonimización y las garantías de tratamiento ofrecidas por el sistema (Al-Kfairy et al., 2024). Esta dimensión traduce las obligaciones de confidencialidad y protección de datos al contexto de la IA generativa.

3. Trazabilidad documental mínima

La tercera dimensión es la trazabilidad. Todo uso relevante de IA generativa debe generar un registro mínimo que permita reconstruir el proceso de generación (Li et al., 2025). Este registro incluye identificación del sistema, entradas, salida generada, intervención humana y destino del producto final.

La intensidad del registro debe ser proporcional al riesgo del uso. En tareas de bajo impacto puede ser simplificada, mientras que en usos de alto impacto debe formar parte del expediente del producto jurídico (Li et al., 2025). La trazabilidad permite sostener auditoría, control y eventual atribución de responsabilidad.

4. Supervisión humana significativa

La cuarta dimensión es la supervisión humana, entendida como un deber reforzado de validación. Ningún contenido generado debe incorporarse a un producto jurídico sin revisión sustantiva que verifique su corrección normativa, coherencia y adecuación al caso (Tarun et al., 2025).

Esta supervisión implica capacidad real de corregir o descartar la salida del sistema. Su función es evitar que la plausibilidad lingüística del contenido sustituya el juicio jurídico, manteniendo la responsabilidad profesional en el centro del proceso (Hahn et al., 2021).

5. Gestión de riesgo

La quinta dimensión consiste en la gestión del riesgo. La gobernanza debe distinguir entre usos permitidos, restringidos y no admisibles, en función del impacto potencial sobre derechos y de la naturaleza de la tarea (OECD, 2023).

Se consideran usos restringidos aquellos que implican datos sensibles o decisiones con efectos jurídicos relevantes. Los usos no admisibles incluyen la incorporación de información protegida en entornos no controlados o el uso de contenido no verificado como base de decisiones jurídicas (Mantelero, 2024). Esta clasificación permite prevenir daños sin impedir usos legítimos.

6. Responsabilidad documentable

La sexta dimensión es la responsabilidad documentable. El uso de IA generativa debe permitir identificar quién intervino en cada fase del proceso: quién autorizó el uso, quién ingresó datos, quién validó el contenido y quién adoptó el producto final (Bauer et al., 2023).

Esta estructura evita la responsabilidad difusa y permite articular *accountability* en entornos híbridos (OECD, 2023). Además, facilita la revisión institucional y el aprendizaje organizacional a partir de incidentes o errores. La Tabla 2 presenta la matriz jurídica operativa propuesta, articulando cada dimensión con su exigencia mínima, indicador operativo, evidencia documental, responsable principal y ejemplo de aplicación práctica.

Tabla 2. Matriz jurídica operativa para el uso legítimo de IA generativa en el ecosistema jurídico

Dimensión	Exigencia jurídica mínima	Indicador operativo	Evidencia documental	Responsable principal	Ejemplo de aplicación
Delimitación del caso de uso	Definir si el uso es exploratorio, auxiliar o de alto impacto jurídico	Ficha de clasificación previa del uso	Registro del caso de uso, finalidad y nivel de impacto	Organización, despacho, tribunal o responsable del área jurídica	Distinguir entre usar IA para lluvia de ideas, preparar un resumen interno o elaborar un borrador que será incorporado a un escrito jurídico
Gobernanza de datos	Aplicar minimización, confidencialidad y control de acceso	Verificación previa de datos personales, sensibles, estratégicos o confidenciales	Checklist de datos introducidos, anonimización aplicada y base de tratamiento	Usuario jurídico y responsable de protección de datos o cumplimiento	Evitar introducir nombres, expedientes completos, datos sensibles o información protegida por secreto profesional en sistemas no controlados
Trazabilidad documental	Permitir la reconstrucción verificable de la interacción con el sistema	Registro de herramienta, versión, fecha, prompt, salida generada y modificaciones humanas	Bitácora de uso, anexo interno de trazabilidad o registro de expediente	Usuario jurídico que opera la herramienta	Guardar el prompt y la respuesta cuando la salida sea utilizada para preparar un contrato, dictamen, demanda o informe jurídico
Supervisión humana significativa	Validar jurídicamente el contenido antes de incorporarlo a un producto final	Revisión sustantiva de normas, jurisprudencia, hechos, citas y razonamiento	Versión revisada, comentarios de validación y constancia de aprobación	Profesional jurídico responsable de la validación	Verificar que las citas legales y jurisprudenciales existan, sean pertinentes y estén correctamente interpretadas
Gestión del riesgo	Ajustar controles según impacto jurídico y posible afectación de derechos	Clasificación del uso como permitido, restringido o no admisible	Matriz de riesgo, decisión de autorización y condiciones de uso	Comité, socio responsable, juez, coordinador jurídico o área de cumplimiento	Restringir el uso de IA para decisiones que afecten derechos, estrategias procesales o información confidencial sin revisión reforzada
Responsabilidad documentable	Identificar quién autorizó, operó, revisó y aprobó el uso de IA	Asignación de roles por fase del proceso	Registro de responsables, firma, constancia de revisión o aprobación final	Institución y profesional jurídico validador	Determinar quién responde si un escrito incorpora una cita inexistente, una interpretación errónea o información no verificada generada por IA

En conjunto, esta matriz establece un estándar jurídico-operativo de uso legítimo que permite trasladar principios generales, como explicabilidad, transparencia, supervisión humana y responsabilidad, a criterios verificables dentro de la práctica jurídica (Floridi y Cowls, 2019). Su valor no radica únicamente en ordenar categorías conceptuales, sino en indicar qué debe documentarse, quién debe intervenir, qué evidencia debe conservarse y cómo debe graduarse el control según el riesgo del caso de uso. Por ello, la matriz puede funcionar como instrumento de cumplimiento interno, guía de diligencia profesional y soporte de auditoría en organizaciones jurídicas que integren IA generativa en procesos de redacción, análisis, revisión documental o apoyo a la toma de decisiones.

Para ejemplificar su aplicación práctica, puede considerarse el caso de un despacho jurídico que utiliza IA generativa para preparar un primer borrador de contestación de demanda. Bajo un enfoque meramente conceptual, bastaría con afirmar que la herramienta debe ser transparente o explicable. En cambio, bajo la matriz propuesta, el despacho tendría que realizar al menos seis operaciones verificables: clasificar el uso como apoyo profesional de impacto jurídico, excluir o anonimizar datos innecesarios, conservar el *prompt* y la salida generada, registrar las modificaciones introducidas por el abogado, verificar la existencia y pertinencia de normas o precedentes citados, y documentar quién aprobó la versión final. Esta secuencia convierte la gobernanza en evidencia, porque no solo declara que existió supervisión humana, sino que permite reconstruir cómo se realizó, sobre qué contenido recayó y quién asumió la validación final.

El mismo razonamiento puede trasladarse a un tribunal, una clínica jurídica universitaria o un área legal corporativa. En estos contextos, la IA generativa podría utilizarse legítimamente para resumir documentos extensos, elaborar índices preliminares, comparar argumentos o preparar versiones iniciales de textos jurídicos. Sin embargo, el uso se volvería problemático si la salida se incorporara directamente a una resolución, demanda, contrato o dictamen sin trazabilidad ni revisión sustantiva. La matriz permite distinguir ambos escenarios: no prohíbe el uso de IA generativa, pero condiciona su legitimidad a que el proceso sea documentado, supervisado, proporcional al riesgo y jurídicamente imputable.

VII. Objeciones doctrinales y límites prácticos de la matriz propuesta

La trazabilidad no debe presentarse como una solución exenta de tensiones. Una primera objeción sostiene que los deberes intensivos de documentación pueden elevar los costes de implementación y generar cargas administrativas desproporcionadas, especialmente en despachos pequeños, clínicas jurídicas, organizaciones públicas con baja madurez tecnológica o áreas legales que operan con recursos limitados. Esta crítica es relevante porque la transparencia, cuando se traslada a procedimientos concretos, puede producir fricciones entre el ideal normativo de control, la capacidad técnica disponible y los recursos organizacionales necesarios para sostener registros, auditorías y revisiones internas (Felzmann *et al*., 2020). Sin embargo, esta objeción no invalida la trazabilidad, sino que obliga a diseñarla bajo un criterio de proporcionalidad. No todos los

usos de IA generativa requieren el mismo nivel de registro, pues una lluvia de ideas preliminar no debe documentarse con la misma intensidad que un borrador incorporado a una demanda, contrato, dictamen o resolución. Por ello, la matriz propuesta adopta una lógica gradual: a mayor impacto jurídico, mayor exigencia de trazabilidad, supervisión y responsabilidad documentable.

Una segunda objeción se refiere a la posible tensión entre trazabilidad, secreto profesional, confidencialidad y protección de secretos empresariales. Documentar prompts, entradas, salidas, versiones del sistema y procesos de revisión podría generar nuevos riesgos si la organización conserva información sensible sin controles adecuados o si la trazabilidad se interpreta como obligación de revelar públicamente información estratégica del caso, datos de clientes o detalles técnicos protegidos por propiedad intelectual. La literatura sobre transparencia algorítmica ha advertido que los sistemas de IA suelen estar protegidos mediante secreto empresarial y que esa protección puede reforzar la opacidad, especialmente cuando se invoca para limitar el escrutinio externo del sistema (Foss-Solbrekk, 2021). Por tanto, la trazabilidad jurídica no debe confundirse con publicidad total ni con apertura irrestricta del sistema, sino con documentación interna controlada, proporcional y accesible solo para fines legítimos de revisión, auditoría, cumplimiento o defensa. Esta distinción permite compatibilizar trazabilidad, secreto profesional y protección de datos, evitando que el remedio de la transparencia produzca nuevos riesgos de exposición indebida.

Una tercera objeción apunta al riesgo de formalismo documental. Una institución podría conservar bitácoras de uso, prompts y respuestas generadas por IA sin realizar una verdadera revisión sustantiva del contenido. En ese caso, la trazabilidad se convertiría en un expediente burocrático de cumplimiento aparente, pero no en una garantía real de diligencia profesional. Este riesgo es particularmente grave en el ámbito jurídico, porque la existencia de un registro no demuestra por sí misma que se hayan verificado normas, jurisprudencia, hechos, citas, argumentos o consecuencias jurídicas. La experiencia reciente muestra que el uso acrítico de IA generativa en escritos jurídicos puede producir consecuencias disciplinarias o procesales cuando se incorporan citas inexistentes o fundamentos no verificados, como ocurrió en Mata v. Avianca, Inc., donde se sancionó la presentación de precedentes inexistentes generados mediante ChatGPT (*Mata v. Avianca, Inc.*, 2023). En consecuencia, la matriz propuesta separa dos exigencias que no deben confundirse: la trazabilidad documental del proceso y la supervisión humana significativa del resultado.

Una cuarta objeción se relaciona con la implementación técnica y organizacional. La trazabilidad requiere definir quién registra, dónde se conserva la información, durante cuánto tiempo, bajo qué medidas de seguridad y con qué criterios de acceso. También exige capacitar a los usuarios jurídicos para distinguir entre usos exploratorios, auxiliares y de alto impacto, así como establecer reglas institucionales sobre herramientas permitidas, datos prohibidos, validación de resultados y escalamiento de incidentes. Las guías éticas recientes sobre IA generativa en la profesión jurídica insisten en que el uso de estas herramientas debe conectarse con deberes de competencia, confidencialidad, comunicación, supervisión y cobro razonable de honorarios, lo que confirma que la gobernanza

no puede reducirse a una exigencia técnica aislada (American Bar Association, 2024). Por ello, la matriz propuesta debe entenderse como una herramienta de madurez progresiva: puede iniciar con registros mínimos y checklists de validación, pero debe evolucionar hacia políticas internas, controles de acceso, auditorías y mecanismos de revisión cuando el uso de IA generativa tenga mayor impacto jurídico.

Estas objeciones permiten precisar el alcance de la propuesta. La trazabilidad defendida en este artículo no equivale a registrar todo, revelar todo o auditar todo con la misma intensidad, sino a conservar la evidencia mínima necesaria para reconstruir usos jurídicamente relevantes de IA generativa. Tampoco sustituye el juicio profesional, sino que lo hace verificable. De este modo, la matriz evita dos extremos igualmente problemáticos: por un lado, la opacidad absoluta basada en la complejidad técnica, el secreto profesional o el secreto empresarial; por otro, una transparencia indiscriminada que aumente costes, exponga información sensible o genere cargas inviables. La solución propuesta es una trazabilidad proporcional, finalista y jurídicamente situada, orientada a permitir control, revisión y atribución de responsabilidad sin sacrificar confidencialidad ni viabilidad operativa. La Tabla 3 sintetiza las principales objeciones doctrinales y límites prácticos identificados, así como la respuesta que ofrece la matriz propuesta desde una lógica de proporcionalidad, control y viabilidad institucional.

Tabla 3. Objeciones doctrinales y respuesta de la matriz propuesta

Objeción o límite práctico	Riesgo jurídico u organizacional	Respuesta de la matriz propuesta
Coste de implementación	La documentación intensiva puede ser inviable para despachos pequeños, clínicas jurídicas o instituciones con baja madurez tecnológica	Aplicar trazabilidad proporcional según el nivel de impacto jurídico del caso de uso
Sobrecarga administrativa	La trazabilidad puede convertirse en una carga burocrática que reduzca eficiencia sin mejorar control real	Utilizar registros mínimos, checklists y evidencias diferenciadas por riesgo
Secreto profesional	Documentar prompts, entradas y salidas puede exponer información confidencial de clientes o estrategias jurídicas	Implementar documentación interna controlada, anonimización, restricciones de acceso y conservación limitada
Secreto empresarial y propiedad intelectual	Los proveedores pueden invocar secreto empresarial para limitar explicaciones técnicas o auditorías externas	Distinguir entre explicabilidad técnica del modelo y trazabilidad jurídica del uso concreto realizado por la organización

Objeción o límite práctico	Riesgo jurídico u organizacional	Respuesta de la matriz propuesta
Formalismo documental	La existencia de registros puede simular cumplimiento sin revisión sustantiva del resultado	Separar trazabilidad documental y supervisión humana significativa como exigencias acumulativas
Dificultad organizacional	La matriz exige roles, capacitación, políticas internas y criterios de conservación de evidencia	Implementar la matriz como herramienta de madurez progresiva y no como obligación uniforme inmediata

VIII. Conclusiones

La incorporación de inteligencia artificial generativa en el ecosistema jurídico no constituye únicamente una innovación instrumental, sino una transformación estructural en la producción, validación y circulación del razonamiento jurídico. Esta transformación exige replantear las condiciones bajo las cuales el uso de estas tecnologías puede considerarse compatible con estándares de diligencia profesional, confidencialidad, protección de datos y responsabilidad institucional.

El artículo ha sostenido que la explicabilidad, aunque necesaria, resulta insuficiente como estándar autónomo de control. La posibilidad de comprender el funcionamiento general de un sistema o de obtener una justificación aparente de sus resultados no garantiza la corrección jurídica del contenido ni permite, por sí sola, reconstruir el proceso de generación o atribuir responsabilidad por su uso. En el ámbito jurídico, donde el lenguaje generado puede incorporarse directamente a productos con efectos normativos, esta limitación adquiere especial relevancia.

Frente a ello, se propuso una lectura integrada basada en una gobernanza orientada a derechos, en la que la explicabilidad se articula con trazabilidad, supervisión humana significativa, gestión del riesgo, protección de datos y responsabilidad documentable. Este enfoque permite desplazar el análisis desde la mera inteligibilidad técnica hacia la posibilidad de someter el uso de IA generativa a condiciones verificables de control, revisión y atribución de deberes.

En particular, la trazabilidad emerge como un elemento central, al permitir reconstruir el proceso de generación, sostener mecanismos de auditoría y ofrecer soporte probatorio en contextos de revisión o controversia. Sin trazabilidad, la explicabilidad carece de valor operativo; con ella, el uso de la IA generativa se vuelve jurídicamente gobernable.

La principal aportación del trabajo consiste en la formulación de una matriz jurídica operativa de uso legítimo que traduce estándares normativos y principios de gobernanza en criterios verificables para la práctica jurídica. Al integrar delimitación del caso de uso, gobernanza de datos, trazabilidad documental, supervisión humana significativa, gestión del riesgo y responsabilidad documentable, la matriz permite identificar qué debe documentarse, quién debe intervenir,

qué evidencia debe conservarse y cómo debe graduarse el control según el impacto jurídico del uso. De este modo, la gobernanza deja de formularse como una declaración general de principios y se convierte en una secuencia de decisiones verificables: clasificar el caso de uso, controlar los datos introducidos, conservar evidencia del proceso, validar sustantivamente el resultado, graduar el riesgo y asignar responsabilidad documentable. Esta estructura diferencia la propuesta de marcos éticos generales, pues ofrece una herramienta aplicable como guía de cumplimiento, parámetro de diligencia profesional y soporte de auditoría para despachos, tribunales, clínicas jurídicas, áreas legales institucionales y entidades públicas.

El análisis también permite reconocer que la trazabilidad no está exenta de límites. Su implementación puede generar costes, cargas administrativas, tensiones con el secreto profesional, dificultades frente al secreto empresarial de los proveedores y riesgos de formalismo documental. Sin embargo, estas objeciones no eliminan su valor jurídico, sino que obligan a diseñarla bajo criterios de proporcionalidad, finalidad y control de acceso. Por ello, la propuesta no defiende una trazabilidad absoluta o indiscriminada, sino una trazabilidad jurídicamente situada: suficiente para reconstruir usos relevantes de IA generativa, pero limitada por la confidencialidad, la protección de datos, la viabilidad organizacional y la intensidad del riesgo.

En definitiva, la legitimidad del uso de inteligencia artificial generativa en el ecosistema jurídico no depende únicamente de la capacidad de explicar sus resultados, sino de la posibilidad de someter su utilización a condiciones verificables de trazabilidad, supervisión, proporcionalidad e imputación jurídica. Solo bajo estas condiciones es posible compatibilizar la innovación tecnológica con la protección de derechos, la confidencialidad, la diligencia profesional y la integridad del ejercicio del Derecho.

Referencias bibliográficas

Abhishek, A., Erickson, L. y Bandopadhyay, T. (2025). Data and AI governance: Promoting equity, ethics, and fairness in large language models (Versión 1). *arXiv*. https://doi.org/10.48550/ARXIV.2508.03970

ACM. (2023). *ACM Code of Ethics and Professional Conduct*. https://www.acm.org/code-of-ethics

Adiguzel, T., Kaya, M. H. y Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, *15*(3), ep429. https://doi.org/10.30935/cedtech/13152

Afroogh, S., Akbari, A., Malone, E., Kargar, M. y Alambeigi, H. (2024). Trust in AI: Progress, Challenges, and Future Directions (Versión 3). *arXiv*. https://doi.org/10.48550/ARXIV.2403.14680

Agbese, M., Mohanani, R., Khan, A. y Abrahamsson, P. (2023). Implementing AI Ethics: Making Sense of the Ethical Requirements. *Proceedings of the*

27th International Conference on Evaluation and Assessment in Software Engineering, 62-71. https://doi.org/10.1145/3593434.3593453

Ahmad, M., Alhalaiqa, F. y Subih, M. (2023). Constructing and testing the psychometrics of an instrument to measure the attitudes, benefits, and threats associated with the use of Artificial Intelligence tools in higher education. *Journal of Applied Learning and Teaching*, *6*(2), 114-120. https://doi.org/10.37074/jalt.2023.6.2.36

Aleem, Y. A. (2026). Socially Responsible AI in the GPT Era. *SMU Law Review*, *78*(3), 563. https://doi.org/10.25172/smulr.78.3.7

Al-Kfairy, M., Mustafa, D. G., Kshetri, N., Insiew, M. y Alfandi, O. (2024). Ethical challenges and solutions of generative AI: An interdisciplinary perspective. *Informatics*, *11*(3). https://doi.org/10.3390/informatics11030058

Allaham, M. y Diakopoulos, N. (2024). Evaluating the Capabilities of LLMs for Supporting Anticipatory Impact Assessment (Versión 2). *arXiv*. https://doi.org/10.48550/ARXIV.2401.18028

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E. y Zayed, T. (2023). Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems*, *11*(7), 351. https://doi.org/10.3390/systems11070351

Altukhi, Z. M. y Pradhan, S. (2025). Systematic Literature Review: Explainable AI Definitions and Challenges in Education (arXiv:2504.02910). *arXiv*. https://doi.org/10.48550/arXiv.2504.02910

Alvarado, R. (2023). AI as an Epistemic Technology. *Science and Engineering Ethics*, *29*(5), 32. https://doi.org/10.1007/s11948-023-00451-3

Ananny, M. y Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973-989. https://doi.org/10.1177/1461444816676645

Atenas, J. (2024). *Considerations to navigate copyright issues in the context of emerging technologies and OER*. https://doi.org/10.13140/RG.2.2.17475.21289

Awasthi, R., Bhattad, A., Ramachandran, S. P., Mishra, S., Khanna, A. K., Cywinski, J. B., Maheshwari, K., Mahapatra, D., Dirosa, I., Cohen, A., Arshad, H., Atreja, A., Alshukaili, A., Vohra, A., Singh, N., Papay, F. A., Atreja, A., Kashyap, R. y Mathur, P. (2025). Human evaluation of large language models in healthcare: Gaps, challenges, and the need for standardization. *NPJ Health Systems*, *2*(1), 40. https://doi.org/10.1038/s44401-025-00043-2

Bach, T. A., Khan, A., Hallock, H., Beltrão, G. y Sousa, S. (2024). A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human-Computer Interaction*, *40*(5), 1251-1266. https://doi.org/10.1080/10447318.2022.2138826

Bandi, A., Adapa, P. V. S. R. y Kuchi, Y. E. V. P. K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input-Output Formats, Evaluation Metrics, and Challenges. *Future Internet*, *15*(8), 260. https://doi.org/10.3390/fi15080260

Bano, M., Zowghi, D., Shea, P. e Ibarra, G. (2023). Investigating Responsible AI for Scientific Research: An Empirical Study (Versión 1). *arXiv*. https://doi.org/10.48550/ARXIV.2312.09561

Barocas, S., Hardt, M. y Narayanan, A. (2018). *Fairness and Machine Learning Limitations and Opportunities*. https://api.semanticscholar.org/CorpusID:113402716

Bathaee, Y. (2018). The artificial intelligence black box and the failure of intent and causation. Harvard Journal of Law & Technology, 31(2), 890-938.

Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., Fischer, M. R. y Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology*, *54*(5), 1222-1245. https://doi.org/10.1111/bjet.13336

Bazanova, E. M., Gorizontova, A. V., Gribova, N. N., Chikake, T. M. y Samosyuk, A. V. (2023). Development and Prospects of National Intelligent System for Testing General Language Competencies Deployed Through Neural Network Solutions. *Vysshee Obrazovanie v Rossii*, *32*(8-9), 147-166. https://doi.org/10.31992/0869-3617-2023-32-8-9-147-166

Bearman, M. y Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, *54*(5), 1160-1173. https://doi.org/10.1111/bjet.13337

Bellamy, R. (2017). Ronald Dworkin, Taking Rights Seriously. En J. T. Levy (Ed.), *The Oxford Handbook of Classics in Contemporary Political Theory*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198717133.013.18

Ben-Michael, E., Greiner, D. J., Huang, M., Imai, K., Jiang, Z. y Shin, S. (2024). Does AI help humans make better decisions? A statistical evaluation framework for experimental and observational studies (Versión 3). *arXiv*. https://doi.org/10.48550/ARXIV.2403.12108

Birkstedt, T., Minkkinen, M., Tandon, A. y Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, *33*(7), 133-167. https://doi.org/10.1108/INTR-01-2022-0042

Breen, S., Ouazzane, K. y Patel, P. (2020). GDPR: Is your consent valid? *Business Information Review*, *37*(1), 19-24. https://doi.org/10.1177/0266382120903254

Casey, B., Farhangi, A. y Vogl, R. (2019). Rethinking explainable machines: The gdpr's 'right to explanation' debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*, *34*(1), 143. https://doi.org/10.15779/Z38M32N986

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, *6*, 1421273. https://doi.org/10.3389/fhumd.2024.1421273

Choudhary, S., Anshumaan, D., Palumbo, N. y Jha, S. (2025). How Not to Detect Prompt Injections with an LLM (Versión 3). *arXiv*. https://doi.org/10.48550/ARXIV.2507.05630

Clark, P., Tafjord, O. y Richardson, K. (2020). Transformers as Soft Reasoners over Language (Versión 2). *arXiv*. https://doi.org/10.48550/ARXIV.2002.05867

Cornejo-Plaza, I. y Cippitani, R. (2023). Ethical and Legal Considerations of Artificial Intelligence in Higher Education: Challenges and Prospects. *Revista de Educacion y Derecho*, (28). https://doi.org/10.1344/REYD2023.28.43935

Cotton, D. R. E., Cotton, P. A. y Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2). https://doi.org/10.1080/14703297.2023.2190148

De Oliveira Fornasier, M. (2021). Legal Education in the 21st Century and the Artificial Intelligence. *Revista Opinião Juridica*, *19*(31), 1-32. https://doi.org/10.12662/2447-6641OJ.V19I31.P1-32.2021

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., De Prado, M. L., Herrera-Viedma, E. y Herrera, F. (2023). Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation (Versión 2). *arXiv*. https://doi.org/10.48550/ARXIV.2305.02231

Feldvari, K., Mi□unovi□, M. y Badurina, B. (2022). Hacking the crisis of democracy: Can media literacy save us from algorithmic modelling of political perception, will and thought? *Vjesnik Bibliotekara Hrvatske*, *65*(2), 23-48. https://doi.org/10.30754/vbh.65.2.971

Felin, T. y Holweg, M. (2024). Theory Is All You Need: AI, Human Cognition, and Causal Reasoning. *Strategy Science*, *9*(4), 346-371. https://doi.org/10.1287/stsc.2024.0189

Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*, *28*(11). https://doi.org/10.5210/fm.v28i11.13346

Ferrer, X., Nuenen, T. V., Such, J. M., Cote, M. y Criado, N. (2021). Bias and Discrimination in AI: A Cross-Disciplinary Perspective. *IEEE Technology and Society Magazine*, *40*(2), 72-80. https://doi.org/10.1109/MTS.2021.3056293

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press. https://doi.org/10.1093/oso/9780198883098.001.0001

Floridi, L. y Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P.

y Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., King, T. C. y Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, *26*(3), 1771-1796. https://doi.org/10.1007/s11948-020-00213-5

Gangavarapu, R. (2025). Unveiling the Black Box: Enhancing AI/ML Model Explainability for Transparency and Trust. En R. Gangavarapu, *Mastering AI Governance* (pp. 87-96). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-93681-4_9

García López, I. M., Ramírez Montoya, M. S. y Molina Espinosa, J. M. (2024). Design and challenges of open large language model frameworks (Open LLM): A systematic literature mapping. *ICERI 2024 Proceedings*. https://doi.org/10.21125/iceri.2024

Ghimire, A. y Edwards, J. (2024). From Guidelines to Governance: A Study of AI Policies in Education (arXiv:2403.15601). *arXiv*. https://doi.org/10.48550/arXiv.2403.15601

Golovianko, M., Gryshko, S., Terziyan, V. y Tuunanen, T. (2023). Responsible cognitive digital clones as decision-makers: a design science research study. *European Journal of Information Systems*, *32*(5), 879-901. https://doi.org/10.1080/0960085X.2022.2073278

Grimmelikhuijsen, S. y Meijer, A. (2022). Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. *Perspectives on Public Management and Governance*, *5*(3), 232-242. https://doi.org/10.1093/ppmgov/gvac008

Grünewald, E. y Pallas, F. (2021). TILT: A GDPR-Aligned Transparency Information Language and Toolkit for Practical Privacy Engineering. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 636-646. https://doi.org/10.1145/3442188.3445925

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. y Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 1-42. https://doi.org/10.1145/3236009

Hahn, M., Jurafsky, D. y Futrell, R. (2021). Sensitivity as a Complexity Measure for Sequence Classification Tasks. *Transactions of the Association for Computational Linguistics*, *9*, 891-908. https://doi.org/10.1162/tacl_a_00403

Hasibuan, R. H., Rawung, A. J. y Wowiling, F. J. (2024). Indonesian Law and Artificial Intelligence: Balancing Accountability, Ethics, and Innovation. *Jurnal Penelitian Hukum De Jure*, *24*(2), 121. https://doi.org/10.30641/dejure.2024.V24.121-132

Hassoulas, A., Powell, N., Roberts, L., Umla-Runge, K., Gray, L. y Coffey, M. J. (2023). Investigating marker accuracy in differentiating between university scripts written by students and those produced using ChatGPT. *Journal of Applied Learning and Teaching*, *6*(2), 71-77. Scopus. https://doi.org/10.37074/jalt.2023.6.2.13

Head, C. B., Jasper, P., McConnachie, M., Raftree, L. y Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. *New Directions for Evaluation*, (178-179). https://doi.org/10.1002/ev.20556

Herrera, F. y Calderón, R. (2025). Opacity as a Feature, Not a Flaw: The LoBOX Governance Ethic for Role-Sensitive Explainability and Institutional Trust in AI (Versión 1). *arXiv*. https://doi.org/10.48550/ARXIV.2505.20304

Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S. y Zimmermann, A. (2022). Knowledge Graphs. *ACM Computing Surveys*, *54*(4), 1-37. https://doi.org/10.1145/3447772

Kelly, A., Sullivan, M. y Strampel, K. (2023). Generative artificial intelligence: University student awareness, experience, and confidence in use across disciplines. *Journal of University Teaching and Learning Practice*, *20*(6). https://doi.org/10.53761/1.20.6.12

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, *24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Kreps, S., & Kriner, D. (2023). How AI Threatens Democracy. *Journal of Democracy*, *34*(4), 122-131. https://doi.org/10.1353/jod.2023.a907693

Leung, T. I., Sagar, A., Shroff, S., & Henry, T. L. (2023). Can AI Mitigate Bias in Writing Letters of Recommendation? *JMIR Medical Education*, *9*. https://doi.org/10.2196/51494

Li, Z., Zhu, H., Lu, Z., Xiao, Z. y Yin, M. (2025). From Text to Trust: Empowering AI-assisted Decision Making with Adaptive LLM-powered Analysis. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1-18. https://doi.org/10.1145/3706598.3713133

Mantelero, A. (2024). Retos y regulación de la Inteligencia Artificial: La toma de decisiones en los asuntos públicos y la administración de justicia. En P. Valcárcel Fernández y F. L. Hernández González (Coords.), *El Derecho Administrativo en la era de la inteligencia artificial: actas del XVIII Congreso de la Asociación Española de Profesores de Derecho Administrativo*. Instituto Nacional de Administración Pública.

Marín, V. I. y Tur, G. (2023). Data privacy in Educational Technology: Findings of a scoping review. *Edutec*, *83*, 7-23. https://doi.org/10.21556/edutec.2023.83.2701

Méndez Carpio, C. R., Arévalo Medranda, S. R., León Segovia, L. S., Parra Guerrero, E. F. y Siguencia Tello, S. P. (2023). Tecnopatías y dependencias: Uso incorrecto de la tecnología. *Killkana Social*, *7*(3), 77-88. https://doi.org/10.26871/killkanasocial.v7i3.1407

Núñez Portilla, J. E. y Guerrero Zambrano, M. F. (2025). Inteligencia artificial en la investigación científica: Una revisión crítica de las herramientas de análisis bibliográfico y de datos en la educación superior: Artificial intelligence

in scientific research: a critical review of bibliographic and data analysis tools in higher education. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, *6*(4). https://doi.org/10.56712/latam.v6i4.4537

OECD. (2023). *Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI*. https://doi.org/10.1787/2448f04b-en

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). GPT-4 Technical Report (Versión 6). *arXiv*. https://doi.org/10.48550/ARXIV.2303.08774

Rambachan, A. (2024). Identifying Prediction Mistakes in Observational Data. *The Quarterly Journal of Economics*, *139*(3), 1665-1711. https://doi.org/10.1093/qje/qjae013

Schneegans, S., Lewis, J. y Straza, T. (2021). The race against time for smarter development. *Unesco Science Report: The race against time for smarter development* (30-78). Unesco.

Silva, A. A. (2022). Gobernanza, poder y autonomía universitaria en la era de la innovación. *Perfiles Educativos*, *44*(178). https://doi.org/10.22201/iisue.24486167e.2022.178.60735

Stanford Accelerator for Learning. (2023). *Generative AI*. Stanford University. https://acceleratelearning.stanford.edu/funding/generative-ai/

Stanford Law School. (2023). AI & Access to Justice. *Stanford Center for Legal Informatics*. https://justiceinnovation.law.stanford.edu/projects/ai-access-to-justice/

Tarun, B., Du, H., Kannan, D. y Gehringer, E. F. (2025). Human-in-the-Loop Systems for Adaptive Learning Using Generative AI (arXiv:2508.11062). *arXiv*. https://doi.org/10.48550/arXiv.2508.11062

The EU AI Act. (2025). Key issue. Risk-based approach. *The EU AI Act*. https://www.euaiact.com/key-issue/3?utm=

Theis, S., Jentzsch, S., Deligiannaki, F., Berro, C., Raulf, A. P. y Bruder, C. (2023). Requirements for Explainability and Acceptance of Artificial Intelligence in Collaborative Work. En H. Degen y S. Ntoa (Eds.), *Artificial Intelligence in HCI* (vol. 14050, pp. 355-380). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35891-3_22

Trikoili, A., Georgiou, D., Pappa, C. I. y Pittich, D. (2025). Critical Thinking Assessment in Higher Education: A Mixed-Methods Comparative Analysis of AI and Human Evaluator. *International Journal of Human-Computer Interaction*, *41*(24), 1-14. https://doi.org/10.1080/10447318.2025.2499164

Unión Europea. (2025). *Article 50: Transparency obligations for providers and deployers of certain AI systems*. https://artificialintelligenceact.eu/article/50/

Veale, M. y Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of

the proposed approach. *Computer Law Review International*, *22*(4), 97-112. https://doi.org/10.9785/cri-2021-220402

Zhang, H., Kung, P.-N., Yoshida, M., Broeck, G. V. Den y Peng, N. (2024). Adaptable Logical Control for Large Language Models (Versión 2). *arXiv*. https://doi.org/10.48550/ARXIV.2406.13892

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 252-267

LA TRASCENDENTAL CONDENA A META Y YOUTUBE EN LOS ÁNGELES POR LA CONFIGURACIÓN DE SUS PATRONES OSCUROS[1]

THE LANDMARK RULING AGAINST META AND YOUTUBE IN LOS ANGELES FOR THE DESIGN OF THEIR DARK PATTERNS

Diego Fierro Rodríguez

Letrado de la Administración de Justicia en el Ministerio de Justicia (España)

1 Este trabajo se enmarca en el Proyecto I+D+i de generación de conocimiento y fortalecimiento científico y tecnológico, titulado: «Claves de una Justicia resiliente en plena transformación», (IP. Sonia Calaza), del Ministerio de Ciencia e Innovación, con REF PID2024-155197OB-I00 y en la «Red de investigación»: «Alianzas estratégicas de la Justicia: Educación, Igualdad e Inclusividad» (RED2024-153961-T), Coord. Sonia Calaza, Programa Estatal de Transferencia y Colaboración del Ministerio de Ciencia, Innovación y Universidades, Plan Estatal de Investigación Científica, Técnica y de Innovación 2024-2027.

Resumen

El presente estudio analiza el veredicto emitido el 25 de marzo de 2026 por el Tribunal Superior del Condado de Los Ángeles, que condenó a Meta Platforms y a YouTube por negligencia en el diseño de sus plataformas, causando adicción y daños psicológicos a una menor de edad. El fallo, que otorgó 6 millones de dólares en concepto de daños, representa un punto de inflexión en la jurisprudencia estadounidense al trasladar el foco de la responsabilidad desde el contenido generado por usuarios hacia la arquitectura misma del producto digital. Se examinan las implicaciones de esta sentencia en el marco de la Sección 230 de la Ley de Decencia en las Comunicaciones, la analogía con la litigación contra la industria tabacalera, y la proyección internacional del caso en relación con el Reglamento General de Protección de Datos, el Reglamento de Servicios Digitales y el Reglamento de Inteligencia Artificial de la Unión Europea.

Palabras clave

Responsabilidad algorítmica, diseño adictivo, protección de menores, gobernanza digital, patrones oscuros

Abstract

This study analyzes the verdict issued on March 25, 2026, by the Superior Court of Los Angeles County, which found Meta Platforms and YouTube negligent in the design of their platforms, causing addiction and psychological harm to a minor. The ruling, awarding $6 million in damages, represents a turning point in American jurisprudence by shifting the focus of liability from user-generated content to the very architecture of the digital product. The implications of this judgment are examined within the framework of Section 230 of the Communications Decency Act, the analogy with tobacco litigation, and the international projection of the case in relation to the General Data Protection Regulation, the Digital Services Act, and the Artificial Intelligence Regulation of the European Union.

Keywords

Algorithmic accountability, addictive design, child protection, digital governance, dark patterns

I. Introducción al desplazamiento del eje de responsabilidad del contenido al diseño

El 25 de marzo de 2026, el Tribunal Superior del Condado de Los Ángeles dictó un veredicto de particular trascendencia en el litigio iniciado por KGM, una menor de edad, contra Meta Platforms, Inc. y YouTube, LLC. El jurado declaró negligentes a ambas empresas y fijó una indemnización de seis millones de dólares —4,2 millones atribuidos a Meta y 1,8 millones a YouTube— en concepto de daños compensatorios por ansiedad severa, dismorfia corporal y pensamientos

de autolesión derivados del uso compulsivo de sus plataformas desde los seis años de edad. Posteriormente, el jurado duplicó la cuantía mediante daños punitivos de tres millones de dólares adicionales.

La relevancia jurídica de esta resolución no reside primordialmente en su cuantía económica, sino en el desplazamiento conceptual que opera respecto del régimen de responsabilidad de las plataformas digitales. Durante casi tres décadas, el eje de la discusión jurídica en Estados Unidos gravitó en torno a la Sección 230 de la Ley de Decencia en las Comunicaciones de 1996 (47 USC § 230), que otorga a los proveedores de servicios informáticos una inmunidad casi absoluta respecto del contenido generado por terceros. No obstante, los abogados de la demandante no cuestionaron la aplicabilidad de dicha norma al contenido publicado por usuarios ajenos, sino que caracterizaron a las plataformas como fabricantes de un producto defectuoso -su arquitectura algorítmica, sus decisiones de ingeniería y sus patrones de interacción- cuyo diseño deliberado genera dependencia en menores de edad. Al trasladar el litigio al terreno de la responsabilidad por producto (*products liability*), lograron eludir el escudo de la Sección 230, que no ofrece cobertura en materia de diseño de producto.

Esta estrategia procesal abre un territorio jurídico hasta ahora inexplorado: el sometimiento del diseño de *software* al mismo escrutinio que la ingeniería de puentes o la fabricación de productos farmacéuticos. La cuestión central que plantea el veredicto es si la empresa tecnológica responde civilmente por la forma en que ha diseñado deliberadamente su producto para alterar el comportamiento humano, especialmente cuando dicho diseño incide sobre menores de edad cuya capacidad de autodeterminación se encuentra en desarrollo.

II. El marco probatorio: documentación interna como elemento determinante de la negligencia

Durante la fase de prueba del juicio, los abogados de KGM lograron la admisión como prueba documental de correos electrónicos, memorandos de reuniones ejecutivas y presentaciones internas de Meta e Instagram que nunca habían sido destinadas a la esfera pública. Dicha documentación revelaba que los propios empleados y directivos de la empresa discutían abiertamente los efectos negativos de sus productos en la salud mental de los adolescentes. Entre los documentos presentados figuraban estudios internos que documentaban cómo Instagram afectaba particularmente a las niñas jóvenes, exacerbando problemas de imagen corporal y ansiedad, así como conversaciones en las que ingenieros reconocían que determinadas funcionalidades habían sido diseñadas específicamente para aumentar el «tiempo de pantalla» (*screen time*) sin consideración por el bienestar del usuario.

La analogía que los abogados de la demandante establecieron con la litigación contra la industria tabacalera del siglo XX resulta ilustrativa desde una perspectiva estructural. El Gran Acuerdo de Resolución Tabacalera de 1998, que obligó a las empresas del sector a pagar más de 206.000 millones de dólares en un plazo de 25 años, demostró que cuando el diseño del producto y la ocultación de riesgos se combinan con daño a la salud pública, incluso las industrias más

poderosas pueden verse forzadas a rendir cuentas. Los paralelismos estructurales entre ambos supuestos son notables: las plataformas tecnológicas disponen de documentos internos que evidencian su conocimiento de los efectos de sus productos sobre los adolescentes, han financiado estudios propios que minimizan los riesgos y han negado sistemáticamente la causalidad entre el uso de sus plataformas y el deterioro de la salud mental juvenil.

Desde una perspectiva probatoria, la revelación de documentación interna constituye el elemento subjetivo determinante para acreditar la negligencia. La doctrina de Sánchez Frías (2025), en su tesis doctoral sobre prácticas comerciales personalizadas mediante sistemas de inteligencia artificial, sostiene que la explotación de datos sensibles y factores de vulnerabilidad constituye una práctica desleal independientemente del contenido transmitido, lo cual respalda la tesis de que la responsabilidad puede fundarse en el diseño algorítmico mismo. Por su parte, Berrocal Lanzarot (2025) ha anticipado en su estudio sobre la protección jurídica de los menores ante el reto de la inteligencia artificial que los menores no pueden consentir la explotación de sus vulnerabilidades por sistemas automatizados, lo cual resulta coherente con el razonamiento del jurado de Los Ángeles. España (2025), por su parte, ha denominado «economía invisible de la atención» a la mercantilización de la capacidad de dirigir voluntariamente el foco cognitivo, concepto que subyace al modelo de negocio cuestionado en el litigio.

III. La arquitectura de la adicción: análisis de los mecanismos de diseño cuestionados

El análisis forense de los mecanismos de diseño desplegado durante el juicio no se limitó a acusaciones genéricas, sino que desmenuzó tres funcionalidades específicas cuya combinación genera un entorno de uso compulsivo.

En primer lugar, el desplazamiento infinito (*infinite scroll*). Esta característica elimina los puntos de decisión naturales que permitirían al usuario evaluar si desea continuar interactuando con la plataforma. La ausencia de un final estructural en el flujo de contenido suprime la señal de saciedad que, en condiciones normales, activaría la deliberación consciente sobre la continuación del uso. La relevancia jurídica de este diseño radica en que no constituye un accidente técnico, sino una decisión de ingeniería documentada en patentes corporativas y comunicaciones internas donde ejecutivos discuten explícitamente estrategias para aumentar el «tiempo de permanencia» (*dwell time*).

En segundo lugar, los sistemas de recomendación algorítmica. Mientras un medio de comunicación tradicional selecciona contenidos basándose en criterios editoriales, los algoritmos de Meta y YouTube están optimizados exclusivamente para maximizar el *engagement*. El sistema no evalúa qué necesita el usuario, sino únicamente qué le mantendrá más tiempo interactuando con la plataforma. Cuando este sistema se aplica a menores de edad, el resultado es una selección automatizada de contenidos que explota inseguridades, fomenta comparaciones sociales destructivas y amplifica material que genera reacciones emocionales intensas. Martínez Rebollo (2018) ya había señalado que el diseño de las interfaces

digitales incorpora mecanismos de refuerzo intermitente equiparables a los de las máquinas tragamonedas, lo cual resulta particularmente preocupante cuando se aplica a usuarios cuyo sistema de recompensa cerebral se encuentra en pleno desarrollo, como ha documentado Molina Ciudad (2024) en relación con la vulnerabilidad neurobiológica específica de los adolescentes ante estímulos de recompensa social inmediata.

En tercer lugar, la reproducción automática. Al eliminar la fricción deliberada —ese intervalo donde el usuario debería tomar una decisión activa para continuar— el sistema suprime la agencia humana. Sin fricción, no hay decisión consciente; sin decisión consciente, no hay límite autoimpuesto. El contenido continúa de forma automática, atrapando al usuario en un ciclo de estimulación intermitente que replica los principios operativos de las máquinas tragamonedas. Herrero Báguena, Sanz Blas y Zhelyazkova Buzova (2022) identificaron que la falta de fricción en el diseño constituye uno de los antecedentes estructurales de la adicción a redes sociales, mientras que Rubio Guzmán, Hernández Echegaray y Pérez Viejo (2025) subrayaron que la intervención en adicciones requiere comprender que el objeto de la adicción no es la tecnología en sí, sino la arquitectura relacional que ésta impone.

El testimonio de KGM durante el juicio ilustró la progresiva pérdida de control: lo que comenzó como expresión personal se transformó en compulsión, de modo que ya no decidía cuándo usar la aplicación, sino que la aplicación decidía por ella mediante notificaciones diseñadas para generar ansiedad de privación y un flujo interminable que nunca le permitía sentirse «completa». Sus padres documentaron el proceso, buscaron ayuda médica, establecieron la conexión causal y promovieron la acción judicial.

IV. Los daños punitivos: función disuasoria y precedente normativo

Uno de los momentos procesalmente más significativos del juicio tuvo lugar durante la fase de argumentación sobre daños punitivos. El letrado de KGM, Mark Lanier, empleó una metáfora visual —un tarro de caramelos donde cada pieza representaba mil millones de dólares del valor de mercado de las empresas demandadas— para ilustrar al jurado que la responsabilidad civil solo resulta efectiva cuando la sanción produce un efecto disuasorio real. Si la sanción es económicamente insignificante para la empresa, se convierte en un mero costo operativo de hacer negocios.

Los abogados de Meta argumentaron que la empresa ya estaba implementando cambios para proteger a los usuarios jóvenes, invocando el clásico argumento de mejora continua para evitar la imposición de daños punitivos. El abogado de YouTube expresó personalmente su pesar a KGM en la sala. La réplica de Lanier fue inmediata: la disculpa de un letrado no equivale a la rendición de cuentas de la empresa.

El jurado deliberó menos de una hora sobre los daños punitivos y fijó finalmente tres millones de dólares, duplicando la indemnización por daños compensatorios. El objetivo no fue imponer una suma destructora, sino establecer un

precedente de que el diseño algorítmico negligente genera consecuencias jurídicas reales. Como señala González-Orús Charro (2025), sin daños punitivos el derecho se convierte en papel mojado cuando se enfrenta a corporaciones cuyos recursos económicos diluyen la eficacia de las sanciones compensatorias. Plaza Penadés (2023) añade que la opacidad diseñada en las plataformas dificulta extremadamente la prueba individual, por lo que los daños punitivos resultan necesarios para compensar la asimetría probatoria. Mato Pacín (2023) concluye que sin daños punitivos no existe disuasión efectiva en la contratación electrónica mediada por patrones oscuros, tesis que resulta aplicable *mutatis mutandis* al diseño de plataformas sociales.

V. Los acuerdos extrajudiciales de TikTok y Snap: lectura estratégica

Un elemento frecuentemente omitido en el análisis del caso es que TikTok y Snap, originalmente demandadas junto a Meta y YouTube, alcanzaron acuerdos extrajudiciales con la demandante en 2025, antes de que el juicio llegara a veredicto. Ambas empresas tuvieron acceso a la misma documentación interna que se iba a presentar en el juicio y pudieron evaluar sus probabilidades de éxito. Su decisión de llegar a un acuerdo, aunque les evita la etiqueta formal de haber sido declaradas negligentes, constituye en sí misma una forma de comunicación jurídica: preferieron pagar y mantener el litigio en silencio antes que arriesgarse a que un jurado examinara públicamente sus diseños y documentos internos.

Esta dinámica, común en litigios masivos, adquiere implicaciones adicionales en el presente supuesto. El veredicto final solo condenó formalmente a Meta y YouTube, pero sus conclusiones sobre la naturaleza adictiva del diseño de plataformas sociales son aplicables a la dinámica de diseño de todas las empresas del sector. TikTok y Snap quedan en una posición jurídica particularmente vulnerable: carecen de un veredicto favorable que puedan citar, pero tampoco gozan de la «victoria» de haber sido absueltas. Simplemente adquirieron su salida del escrutinio público mediante compensación económica.

Desde una perspectiva económica del derecho, como señalan Herrero Báguena, Sanz Blas y Zhelyazkova Buzova (2022), el coste de llegar a acuerdos extrajudiciales suele ser menor que el de modificar el diseño adictivo, lo cual crea un incentivo perverso que perpetúa las prácticas cuestionadas. Rubio Guzmán, Hernández Echegaray y Pérez Viejo (2025) añaden que los acuerdos extrajudiciales suelen incluir cláusulas de confidencialidad que dificultan la prevención secundaria, al impedir que la información sobre los riesgos del diseño llegue al público. Molina Ciudad (2024) ha encontrado que las plataformas más agresivas en la captación de la atención son precisamente las que más acuerdos alcanzan, lo cual sugiere una correlación inversa entre peligrosidad del diseño y disposición al escrutinio judicial.

VI. El efecto multiplicador: litigios derivados y transformación del riesgo jurídico

El caso de KGM no constituye un litigio aislado, sino lo que en la tradición procesal anglosajona se denomina *bellwether trial* o juicio testigo. Cuando existe una masa crítica de demandas similares —aproximadamente 1600 casos consolidados en California— los tribunales seleccionan uno o varios litigios representativos para someterlos a juicio primero. Aunque el resultado no vincula formalmente a los demás casos, ejerce una fuerza orientadora gravitacional difícil de sobreestimar.

Desde el momento en que un jurado declara negligente a Meta por el diseño de Instagram, los 1600 casos pendientes experimentan una transformación cualitativa: dejan de ser demandas especulativas para convertirse en riesgos concretos con precio. Cada litigio se ha encarecido significativamente, de modo que llevarlos todos a juicio resultaría insostenible para las empresas demandadas. Al mismo tiempo, llegar a acuerdos masivos implicaría reconocer implícitamente la validez de la tesis del diseño negligente que acaba de prosperar.

Meta y Google enfrentan, por tanto, una paradoja estratégica sin salida cómoda. Cualquier estrategia conlleva costos significativos y riesgos incalculables. El veredicto no resuelve los 1600 casos, pero cambia radicalmente la posición negociadora de las partes.

La naturaleza estandarizada de los patrones oscuros convierte cualquier litigio individual en un potencial *class action*, como anticipó Trias (2019). Plaza Penadés (2023) señala que la transparencia forzada por el primer juicio revela patrones comunes a toda la industria, de modo que las conclusiones del veredicto son extrapolables a otras plataformas. González-Orús Charro (2025) añade que las novedades legislativas tienen efectos retroactivos limitados, pero la jurisprudencia sí se aplica a casos anteriores, lo cual amplía el alcance temporal del precedente. Los 1600 casos consolidados en California son, en consecuencia, solo el principio de una potencial oleada litigiosa.

VII. La respuesta europea: convergencia entre regulación preventiva y responsabilidad *ex post*

Mientras el jurado de Los Ángeles deliberaba, las autoridades europeas venían construyendo desde hacía años un edificio normativo preventivo. El Reglamento (UE) 2016/679, General de Protección de Datos (RGPD), ha ilustrado esta tendencia mediante un historial sancionador significativo contra Meta: entre 2021 y 2025, la empresa acumuló más de 3800 millones de euros en multas, incluyendo la sanción récord de 1.200 millones de euros en 2023 por transferencias masivas de datos a Estados Unidos (Resolución de la AEPD de 12 de mayo de 2023, confirmada en sede de recurso).

Desde el 2 de febrero de 2025, el Reglamento (UE) 2024/1689, de Inteligencia Artificial (Reglamento IA), prohíbe expresamente en su artículo 5 las prácticas que manipulen de forma subliminal o engañosa la toma de decisiones, o que exploten vulnerabilidades vinculadas a la edad cuando ello pueda causar daño

significativo. Cuando un jurado norteamericano concluye que Meta diseñó plataformas para generar adicción en niños, la cuestión que se plantea es si dicho diseño no encaja precisamente en las prácticas prohibidas por la regulación europea. El desplazamiento infinito, las recomendaciones que explotan inseguridades, la reproducción automática: ¿no constituyen formas de manipulación que operan más allá de la conciencia del usuario, explotando vulnerabilidades vinculadas a la edad?

El Reglamento (UE) 2022/2065, de Servicios Digitales (DSA), ya obliga a las plataformas con más de 45 millones de usuarios en la Unión Europea a realizar evaluaciones de riesgo sobre el bienestar mental, con atención específica a los menores (artículo 34). El Reglamento IA refuerza esta lógica preventiva. La convergencia entre ambos enfoques —el *ex p*ost estadounidense y el *ex ante* europeo— resulta particularmente significativa: durante años, las grandes tecnológicas han explotado la fragmentación jurídica global, pero el veredicto de California sugiere que esta estrategia de arbitraje regulatorio está perdiendo viabilidad.

VIII. El argumento de la autonomía vulnerada: imposibilidad del consentimiento válido

Trasciende el encuadre técnico-normativo una cuestión de fondo que el derecho no puede seguir ignorando. Un menor de trece años que interactúa durante horas con un algoritmo entrenado para maximizar su tiempo de atención explotando sus necesidades de validación social no está tomando una decisión libre e informada. La arquitectura del producto está calculada para explotar las vulnerabilidades del desarrollo cerebral adolescente, de modo que la noción misma de elección informada se convierte en ficción jurídica.

El consentimiento, como institución jurídica, presupone una capacidad de decisión que el diseño adictivo precisamente anula. Cuando la arquitectura del producto está calculada para explotar las vulnerabilidades neurobiológicas y psicológicas de la adolescencia —documentadas por Molina Ciudad (2024) y por Berral Ortiz, Cánovas Espinal, de la Cruz-Campos y Martínez Domingo (2022), quienes demostraron la alta prevalencia de síntomas adictivos en jóvenes—, cualquier pretendida autonomía del usuario se vicia de origen.

Los veredictos de California y Nuevo México (este último, de fecha próxima, en similar sentido) expresan con claridad que las compañías conocían los riesgos, el diseño era intencionado y el daño era previsible. Esto no constituye mala suerte técnica ni efecto colateral inevitable de la innovación, sino, como mínimo, negligencia. Desde una perspectiva que priorice la dignidad humana sobre la rentabilidad corporativa, cabe calificarlo como la subordinación sistemática del bienestar de los más vulnerables a los imperativos del *engagement*.

IX. Analogía con la industria tabacalera: elementos comunes y diferencias estructurales

La analogía con la industria tabacalera, ya esbozada en el epígrafe II, merece un desarrollo más detallado. Durante décadas, las tabacaleras negaron que sus productos causaran adicción, contando con estudios propios, ejércitos de abogados y una industria política que sostenía su posición. Lo que las derribó no fue un solo juicio, sino una acumulación de litigios insostenible financiera, reputacional y políticamente.

Las similitudes estructurales con el sector tecnológico son notables. Ambas industrias disponen de documentos internos que muestran su conocimiento de los efectos dañinos. Ambas financiaron estudios propios que minimizaban los riesgos. Ambas dirigieron sus estrategias hacia poblaciones vulnerables, incluidos los menores. Y ambas contaron con defensas jurídicas poderosas.

No obstante, existen diferencias importantes. El daño del tabaco era físico y relativamente homogéneo; el daño de las redes sociales es psicológico, difuso y mediado por múltiples factores, lo que dificulta establecer causalidades directas. Por otro lado, las redes sociales poseen una capacidad de adaptación tecnológica mucho más rápida: pueden modificar algoritmos e introducir funcionalidades de "bienestar" en cuestión de semanas, creando la apariencia de mejora sin alterar la lógica subyacente de maximización del *engagement*.

La cuestión abierta es si el ciclo se repetirá. ¿Veremos un «Gran Acuerdo» de las redes sociales? ¿Se establecerán fondos masivos de compensación? Es prematuro responder, pero los elementos estructurales están presentes. Calvo Sánchez-Sierra (2024) ha acuñado la expresión «el infinito en la pantalla» para describir cómo los nuevos vínculos adictivos recrean estructuras de dependencia infantil, lo cual sugiere que el daño es sistémico y no susceptible de soluciones meramente tecnológicas superficiales.

X. Las apelaciones y la irreversibilidad de la transparencia forzada

Google ha anunciado su intención de recurrir el veredicto, y los abogados de Meta probablemente harán lo mismo. El proceso de apelación podría extenderse durante meses o años. En cada nivel, las empresas argumentarán que el jurado se equivocó, que la evidencia no soporta la conclusión de negligencia, que la Sección 230 debería haber protegido sus diseños y que establecer esta responsabilidad ahogará la innovación.

Es posible que tengan éxito en alguna instancia. Los tribunales de apelación, compuestos por jueces profesionales, a veces muestran mayor simpatía por los argumentos corporativos. No obstante, existe algo que la apelación no puede deshacer: el hecho de que un jurado de ciudadanos ordinarios, después de escuchar todas las pruebas y argumentos, concluyó que estas empresas actuaron con negligencia. Ese dato ya forma parte del registro judicial y las 1600 demandas pendientes lo conocen.

Incluso si el veredicto fuera revertido en apelación, la información revelada durante el juicio permanece en el dominio público. No se puede «despublicar» lo que ya se ha dicho en una sala de tribunal. La transparencia forzada tiene valor independiente del resultado final, como señala Navarro Guaimares (2024): el desafío educativo fundamental es desarrollar un pensamiento crítico capaz de interrogar no solo los contenidos, sino los sistemas que los generan. El jurado ejerció precisamente ese pensamiento crítico.

XI. El contexto regulatorio estadounidense: impulso legislativo

El veredicto se produce en un contexto de presión regulatoria sin precedentes en el ámbito tecnológico y digital estadounidense. El Congreso lleva varios años debatiendo propuestas legislativas orientadas a abordar el diseño adictivo de las plataformas digitales y la protección efectiva de los menores en entornos *online*. Entre estas iniciativas destaca el proyecto de ley conocido como *Kids Online Safety Act* (KOSA), junto con otras propuestas similares que buscan imponer mayores responsabilidades a las empresas tecnológicas.

No obstante, estas medidas han tropezado repetidamente con obstáculos políticos, desacuerdos partidistas y la complejidad de equilibrar la innovación con la regulación. En este escenario, el veredicto de Los Ángeles introduce un elemento nuevo y decisivo en el debate público: aporta evidencia concreta que puede ser utilizada por legisladores, reguladores y defensores de la protección infantil para reforzar sus posiciones. La discusión deja de centrarse en riesgos potenciales o daños hipotéticos y pasa a apoyarse en un caso empírico específico, en el que un jurado ha determinado que el diseño de una plataforma fue negligente. Este cambio de enfoque añade un sentido de urgencia política y jurídica, al evidenciar que las consecuencias negativas no son meras especulaciones, sino realidades ya constatadas.

Martinez Velencoso (2025) ha propuesto la ley de equidad digital como marco para un entorno más justo, tesis que resulta coherente con el impulso que el veredicto proporciona a las iniciativas legislativas en curso. Campo Yule, Díaz Mage y Ordoñez (2023) aplicaron técnicas de *machine learning* al consumo de sustancias psicoactivas encontrando patrones análogos a los que emplean las plataformas, lo cual refuerza la necesidad de un marco regulatorio que aborde el diseño adictivo con la misma seriedad que el consumo de sustancias. Gervilla García (2011), en su tesis doctoral sobre aplicación de procedimientos *data mining* a conductas desviadas, ya anticipaba que los algoritmos que predicen comportamientos de riesgo pueden diseñar entornos que los potencien, tesis que resulta particularmente pertinente para comprender cómo los sistemas de recomendación algorítmica no solo predicen sino que amplifican conductas adictivas en usuarios vulnerables.

XII. Vectores de impacto del veredicto: síntesis sistemática

Del análisis desarrollado a lo largo de los epígrafes precedentes se derivan los siguientes vectores de impacto que este veredicto establece:

Primero, la responsabilidad por diseño: el producto —no solo el contenido— puede ser defectuoso y generar responsabilidad civil, lo cual abre la puerta a una litigación masiva basada en la responsabilidad por producto defectuoso, un terreno jurídico donde la inmunidad de la Sección 230 no opera.

Segundo, la erosión del escudo de la Sección 230: las empresas ya no pueden esconderse tras el argumento de que «solo proporcionan la plataforma», pues el diseño algorítmico constituye una conducta activa sujeta a escrutinio.

Tercero, el efecto multiplicador sobre los aproximadamente 1600 casos pendientes: el resultado orienta y presiona la resolución de estos litigios, transformando demandas especulativas en riesgos concretos con precio.

Cuarto, el impulso regulatorio: el veredicto refuerza la posición de los legisladores que impulsan normas específicas sobre diseño adictivo, aportando evidencia empírica a un debate antes dominado por especulaciones.

Quinto, la reconfiguración de los incentivos empresariales: si el riesgo legal se desplaza hacia el diseño, las plataformas deberán internalizar el coste de sus decisiones algorítmicas y priorizar arquitecturas menos nocivas, incluso a expensas del crecimiento y la retención.

Sexto, la consolidación de un estándar probatorio: el veredicto perfila qué evidencias —documentación interna, experimentos A/B, métricas de *engagement*— resultan determinantes para acreditar la relación entre diseño y daño.

Séptimo, la internacionalización del precedente: aunque de origen doméstico estadounidense, su lógica puede permear en otras jurisdicciones, acelerando convergencias regulatorias, especialmente con el marco europeo ya existente.

Octavo, el cambio cultural: se debilita la narrativa de neutralidad tecnológica y se afianza la idea de que el diseño es una forma de conducta con consecuencias jurídicas, sujeta a estándares de diligencia y rendición de cuentas.

De estos vectores se deduce un desplazamiento nítido del eje de responsabilidad: del contenido al diseño, y de la neutralidad proclamada a la rendición de cuentas efectiva. En términos prácticos, emerge un deber de diligencia en la arquitectura del producto que obliga a anticipar riesgos, documentar decisiones y auditar impactos, integrando evaluaciones *ex ante* y controles *ex post*.

XIII. La narrativa jurídica: imposibilidad de control corporativo

La narrativa más conveniente para Meta y Google es que se trata de un caso aislado, de una cantidad de dinero manejable y de un jurado que fue más allá de lo que los tribunales superiores confirmarán. No obstante, la historia de la litigación masiva en Estados Unidos sugiere lo contrario. KGM comenzó a usar YouTube con seis años. Cuando sus padres decidieron demandar, no pensaban en rediseñar internet, sino en la salud de su hija. Pero los efectos de la resolución van mucho más allá de una familia, de una plataforma y de seis millones de dólares.

La cuestión que las democracias occidentales llevan años postergando es quién decide cómo se diseña la atención de una generación. Durante demasiado

tiempo, esta pregunta ha sido respondida exclusivamente por ingenieros de producto en Silicon Valley, trabajando bajo métricas de *engagement*. El veredicto sugiere que existen otros actores legítimos: los usuarios, los padres, los médicos, los educadores, los tribunales y los jurados. La tecnología no es neutral; está diseñada por personas con valores e intereses específicos. Y cuando ese diseño causa daño previsible a los más vulnerables, la responsabilidad jurídica es no solo posible, sino necesaria.

XIV. El comienzo de una era: el algoritmo en el banquillo de los acusados

El mes de marzo de 2026 marca una fecha significativa en la evolución del derecho digital. Por primera vez, el algoritmo se ha sometido realmente a escrutinio judicial, no en Europa con sus sanciones administrativas, sino en California, con el peso específico de un veredicto judicial que declara responsables a Meta y YouTube por fomentar la adicción a redes sociales en menores. No se trata ya de sancionar contenidos ilícitos publicados por terceros, sino de someter a control judicial el propio diseño del producto, el modelo de negocio basado en el *engagement* a cualquier precio y el impacto real que estos sistemas tienen sobre la dignidad y autonomía de los usuarios más vulnerables.

La trascendencia de este fallo reside precisamente en su capacidad para redefinir los términos del debate. Durante demasiado tiempo, la discusión sobre la regulación de las plataformas digitales ha oscilado entre la libertad de expresión y la seguridad, entre la innovación y la protección, como si estos valores fueran inherentemente antagónicos. El veredicto de Los Ángeles sugiere una síntesis diferente: es posible exigir responsabilidad por el diseño sin renunciar a la innovación, proteger a los menores sin censurar la expresión y regular los algoritmos sin destruir la utilidad de las plataformas.

XV. Conclusiones

Del análisis desarrollado a lo largo de los catorce epígrafes precedentes se derivan las siguientes conclusiones, que pretenden sistematizar el significado y el alcance del veredicto del Tribunal Superior del Condado de Los Ángeles.

Primera. El veredicto de 25 de marzo de 2026 consuma un cambio de paradigma en la responsabilidad civil de las plataformas digitales. Al trasladar el foco desde el contenido generado por los usuarios hacia la arquitectura misma del producto —el diseño algorítmico, las decisiones de ingeniería y los patrones de interacción—, la sentencia reconoce que el daño no deriva necesariamente de lo que se dice o muestra en la red, sino de cómo la plataforma ha sido deliberadamente construida para capturar y retener la atención de forma compulsiva. Este giro abre la puerta a una litigación masiva basada en la responsabilidad por producto defectuoso, un terreno jurídico donde la inmunidad de la Sección 230 no opera.

Segunda. La documentación interna presentada durante el juicio ha resultado decisiva para acreditar el elemento subjetivo de la negligencia. Los correos

electrónicos, los memorandos y los estudios internos de Meta demostraron que la empresa conocía los efectos adversos de sus productos sobre la salud mental de los adolescentes y, sin embargo, optó por mantener o incluso intensificar las funcionalidades adictivas. Esta revelación sitúa a la industria tecnológica en una posición análoga a la de la industria tabacalera antes del Gran Acuerdo de 1998, con el agravante de que el daño psicológico y la vulnerabilidad de los menores plantean desafíos probatorios específicos que, sin embargo, el jurado ha sabido resolver aplicando la teoría del diseño negligente.

Tercera. La condena a Meta y YouTube, junto con los acuerdos extrajudiciales alcanzados por TikTok y Snap, confirma que los denominados «patrones oscuros» —desplazamiento infinito, recomendaciones algorítmicas optimizadas para el *engagement*, reproducción automática, notificaciones ansiolíticas— no son meras estrategias comerciales inmorales, sino conductas jurídicamente relevantes que pueden generar responsabilidad civil. Trias (2019) los definió como la «arquitectura de la decepción»; Plaza Penadés (2023) analizó cómo la transparencia en los servicios de intermediación *online* queda sistemáticamente socavada por estos diseños; González-Orús Charro (2025) examinó las novedades legislativas previas al veredicto; Quiroz Ruiz (2025) centró su atención en los patrones oscuros en el tratamiento de datos personales de menores; Ponce Solé (2023) conectó el *nudging* digital con la manipulación y el derecho a una buena administración; Mato Pacín (2023) demostró cómo el consentimiento en la contratación electrónica se vicia cuando median patrones oscuros; y Martínez Rolán (2020) describió cómo el desarrollo web contemporáneo elimina sistemáticamente la libertad real de navegación. La sentencia de Los Ángeles constituye así el primer gran precedente que somete a control judicial la economía de la atención, exigiendo a las plataformas que internalicen los costes sociales de su modelo de negocio.

Cuarta. El efecto multiplicador sobre los aproximadamente 1600 casos pendientes en California es inevitable. El veredicto no los resuelve, pero cambia radicalmente la posición negociadora de las partes. Las empresas demandadas se enfrentan ahora a un dilema estratégico de difícil solución: litigar cada caso individual con el riesgo de acumular veredictos adversos, o llegar a acuerdos masivos que implícitamente reconocerían la validez de la tesis del diseño negligente. En cualquier caso, la litigación secundaria y la presión regulatoria se intensificarán en los próximos años. Cobis y Viloria (2022) señalaron que la alta prevalencia de síntomas adictivos hace enorme el número potencial de afectados; Martínez Rebollo (2018) advirtió que el efecto multiplicador es global; Rubio Guzmán, Hernández Echegaray y Pérez Viejo (2025) alertaron sobre el riesgo de desbordamiento judicial; Molina Ciudad (2024) cuantificó que en España cientos de miles de adolescentes podrían presentar síntomas compatibles; Herrero Báguena, Sanz Blas y Zhelyazkova Buzova (2022) concluyeron que el efecto multiplicador es inevitable; Berral Ortiz, Cánovas Espinal, de la Cruz-Campos y Martínez Domingo (2022) calificaron la adicción a redes como problema de salud pública; Ruíz Méndez, Hernández Ramírez y Flores Morelos (2024) añadieron que los estudiantes universitarios son proclives a demandar tras un precedente exitoso; Clemente Tristán, Guzmán Roa y Salas Blas (2018) relacionaron la adicción

con impulsividad; Basteiro Monje, Robles Fernández, Juarros Basterretxea y Pedrosa (2013) validaron instrumentos de medición que serán usados masivamente; Valencia Ortiz, Garay Ruiz y Cabero Almenara (2020) propusieron medidas formativas que no afectan a los casos ya presentados; y Gauthereau Torres y Godínez Hernández (2021) documentaron la cohorte de jóvenes vulnerables por la pandemia.

Quinta. La convergencia entre la jurisprudencia estadounidense y la regulación europea —Reglamento General de Protección de Datos, Reglamento de Servicios Digitales y especialmente el Reglamento de Inteligencia Artificial de 2025— es un fenómeno de gran significación. Mientras Europa ha optado por un modelo de regulación preventiva *ex ante*, Estados Unidos está construyendo un sistema de responsabilidad *ex post* a través de la litigación civil. El veredicto de Los Ángeles demuestra que ambos enfoques no son contradictorios sino complementarios, y que las grandes tecnológicas ya no pueden explotar la fragmentación jurídica global para eludir estándares mínimos de protección de los derechos fundamentales, en particular de los menores de edad.

Sexta. Desde una perspectiva humanista, el fallo reconoce que el consentimiento prestado por un menor en un entorno diseñado para anular su capacidad de decisión no es un consentimiento válido. La arquitectura adictiva de las plataformas, al explotar las vulnerabilidades neurobiológicas y psicológicas de la adolescencia, vicia de origen cualquier pretendida autonomía del usuario. La sentencia, por tanto, no solo resuelve un caso concreto, sino que sienta las bases para una reconfiguración ética y jurídica del diseño tecnológico: la dignidad humana y la protección de los más vulnerables deben prevalecer sobre la optimización del *engagement* y la monetización de la atención.

Séptima. El veredicto del Tribunal Superior del Condado de Los Ángeles no es el final del proceso, sino su verdadero comienzo. Las apelaciones, los 1600 casos pendientes, las iniciativas legislativas en curso y la evolución de la regulación europea determinarán en los próximos años el alcance definitivo de este cambio de paradigma. Pero lo que ya no puede ser revertido es la narrativa central: el algoritmo ha perdido su inmunidad, y las empresas tecnológicas saben ahora que su diseño puede ser examinado, juzgado y condenado por un jurado de ciudadanos. La responsabilidad algorítmica ha dejado de ser una categoría teórica para convertirse en un riesgo jurídico estructural de primer orden.

Referencias bibliográficas

Basteiro Monje, J., Robles Fernández, A., Juarros Basterretxea, J. y Pedrosa, I. (2013). Adicción a las redes sociales: Creación y validación de un instrumento de medida. *Revista de Investigación y Divulgación en Psicología y Logopedia*, (1), 2-8.

Berral Ortiz, B., Cánovas Espinal, C., De La Cruz-Campos, J. C. y Martínez Domingo, J. A. (2022). Adicción a las redes sociales en jóvenes y adolescentes. En M. Á. Domínguez González, J. C. de la Cruz-Campos, A. Luque de la Rosa y C. Hervás-Gómez (Coords.), *Espacios de aprendizaje e implicaciones prácticas* (pp. 59-67). Dykinson.

Berrocal Lanzarot, A. I. (2025). 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*, (46), 4-32.

Calvo Sánchez-Sierra, M. (2024). El infinito en la pantalla: la infancia, la adolescencia y los nuevos vínculos adictivos en la sociedad virtual. *Revista de psicoanálisis*, (100).

Campo Yule, J. E., Díaz Mage, D. A. y Ordoñez, H. A. (2023). Machine Learning techniques applied to the consumption of illegal psychoactive substances: A systematic mapping. *INGE CUC*, *19*(2), 97-118.

Clemente Tristán, L. A., Guzmán Roa, I. y Salas Blas, E. (2018). Adicción a redes sociales e impulsividad en universitarios de Cusco. *Revista de Psicología*, *8*(1), 13-37.

Cobis, M. y Viloria, E. (2022). Adicción a las redes sociales en adolescentes. *Sistemas Humanos*, *2*(1), 18-33.

España, M. (2025). La economía invisible de la atención, un desafío a la privacidad y a la libertad. *Telos: Cuadernos de comunicación e innovación*, (128), 40-45.

Gauthereau Torres, M. Y. y Godínez Hernández, D. (2021). Adicción a los dispositivos electrónicos y a las redes sociales en tiempos de pandemia. *Milenaria, Ciencia y Arte*, (18), 6-8.

Gervilla García, E. (2011). *Aplicación de procedimientos data mining a conductas desviadas* (Tesis doctoral). Universidad de Santiago de Compostela.

González-Orús Charro, M. (2025). Aspectos básicos sobre los patrones oscuros en las plataformas digitales: novedades legislativas. En M. Dorado Muñoz, M. J. Guerrero Lebrón y L. Alvarado Herrera (Dirs.), *Cuestiones actuales de Derecho mercantil digital* (pp. 625-650). Aranzadi.

Herrero Báguena, B., Sanz Blas, S. y Zhelyazkova Buzova, D. (2022). Antecedentes y consecuencias de la adicción a las redes sociales. En A. Monfort y S. Fernández Lores (Coords.), *Leveraging new business technology for a sustainable economic recovery*. ESIC.

Martínez Rebollo, L. (2018). La adicción a las redes sociales. *Viceversa. UEx & Empresa: La revista para ver, oír, tocar y contar la ciencia*, (91), 4-13.

Martínez Rolán, L. X. (2020). Diseños oscuros en el desarrollo web. Cuando el usuario no es tan libre de elegir una navegación. En D. Caldevilla Domínguez (Coord.), *Unidos por la comunicación: Libro de Actas del XII Congreso Internacional Latina de Comunicación Social*. Latina.

Martinez Velencoso, L. M. (2025). La ley de equidad digital: hacia un entorno digital más justo para los consumidores en la UE. *Revista de Derecho, Empresa y Sociedad*, (25), 19-36.

Mato Pacín, M. N. (2023). Información, consentimiento y patrones oscuros en la contratación electrónica. En J. M. Serrano Cañas, A. Casado Navarro, P. M. González Jiménez, L. M. Miranda Serrano y J. Pagador López (Dirs.), *Contratación mercantil: digitalización y protección del cliente-consumidor* (pp. 235-248). Aranzadi.

Molina Ciudad, M. P. (2024). Adicción a redes sociales en adolescentes. En B. Berral Ortiz, J. A. Martínez Domingo, C. R. Fernández Díaz y J. J. Victoria Maldonado (Coords.), *Investigación para la mejora de las prácticas educativas desde una perspectiva holística* (pp. 345-354). Dykinson.

Navarro Guaimares, J. M. (2024). Pensamiento crítico Vs inteligencia artificial, un desafío para la educación. *Revista Arbitrada: Orinoco, Pensamiento y Praxis*, *14*(2), 17-34.

Plaza Penadés, J. (2023). Transparencia y patrones oscuros en los servicios de intermediación online. En M. E. Cobas Cobiella y R. Guillén Catalán (Dirs.), *Equidad y transparencia en la prestación de servicios* (pp. 97-146). Aranzadi.

Ponce Solé, J. (2023). Derecho, nudging digital y manipulación: patrones oscuros, inteligencia artificial y derecho a una buena administración. *Anuario de la Red Eurolatinoamericana de Buen Gobierno y Buena Administración*, (3).

Quiroz Ruiz, A. I. (2025). Patrones oscuros en el tratamiento de datos personales de menores: desafíos en la era digital. En E. Alcaín Martínez y M. Morillas Fernández (Coords.), *Era digital y personas mayores, con discapacidad y menores: vulnerabilidades y oportunidades* (pp. 891-921). Aranzadi.

Rubio Guzmán, E. M., Hernández Echegaray, A. y Pérez Viejo, J. (2025). Adicción a redes sociales. En J. Páez Gallego (Coord.), *Intervención en adicciones desde el trabajo social* (pp. 211-227). Universitas.

Ruíz Méndez, M., Hernández Ramírez, M. y Flores Morelos, M. M. (2024). Adicción a las redes sociales por estudiantes universitarios. En M. del C. Llorente Cejudo, R. Barragán Sánchez, N. Pérez Rodríguez y L. Martin Párraga (Coords.), *Enseñanza e innovación educativa en el ámbito universitario* (pp. 1630-1635). Dykinson.

Sánchez Frías, I. (2025). *Prácticas comerciales personalizadas mediante sistemas de Inteligencia Artificial. La explotación desleal de datos sensibles y factores de vulnerabilidad* (Tesis doctoral). Universidad de Málaga.

Trias, N. (2019). Patrones oscuros. *Visual: magazine de diseño, creatividad gráfica y comunicación*, *31*(201), 6.

Valencia Ortiz, R., Garay Ruiz, U. y Cabero Almenara, J. (2020). Medidas formativas para la prevención de la adicción a las redes sociales por estudiantes. En E. Colomo Magaña, E. Sánchez Rivas, J. Ruiz Palmero y J. Sánchez Rodríguez (Coords.), *La tecnología como eje del cambio metodológico* (pp. 622-625). UMA Editorial.

Ponencias de investigación ganadoras de los tres primeros premios en la IX edición del Concurso «Valentín Carrascosa» en la categoría de estudiantes, durante el XXVII Congreso de Derecho e Informática de la FIADI celebrado en Lima, Perú en 2025

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 268-278

INTELIGENCIA ARTIFICIAL GENERATIVA EN LA RAMA JUDICIAL: DESAFÍOS PARA LA FUNCIÓN JURISDICCIONAL Y GARANTÍAS DEL DEBIDO PROCESO

GENERATIVE ARTIFICIAL INTELLIGENCE IN THE JUDICIARY: CHALLENGES FOR THE JUDICIAL FUNCTION AND DUE PROCESS GUARANTEES

Shellen Sharay Baez Diaz
Universidad Santo Tomás, seccional Villavicencio

Resumen

El presente artículo analiza los alcances, límites y desafíos del uso de inteligencia artificial generativa en la función judicial en Colombia. A partir del estudio de casos recientes, del marco normativo vigente y de los resultados de la encuesta nacional adelantada por el Consejo Superior de la Judicatura, se examina el potencial de estas herramientas para mejorar la eficiencia judicial, así como las tensiones que genera frente al principio del debido proceso y a la autonomía del juez. Se concluye que, si bien la IA puede apoyar tareas operativas en condiciones reguladas, su intervención no debe sustituir el razonamiento judicial, la valoración probatoria ni la decisión de fondo, funciones que corresponden exclusivamente al operador jurídico. Igualmente, se identifica la necesidad de reducir las brechas digitales regionales y fortalecer los procesos de formación en el uso ético y responsable de estas tecnologías, a fin de garantizar una justicia digital incluyente, eficiente y respetuosa de los derechos fundamentales.

Palabras clave

Inteligencia artificial, Rama Judicial, debido proceso, justicia, automatización

Abstract

This article analyzes the scope, limits, and challenges of using generative artificial intelligence in the judiciary in Colombia. Based on recent case studies, the current regulatory framework, and the results of the national survey conducted by the Superior Council of the Judiciary, the article examines the potential of these tools to improve judicial efficiency, as well as the tensions they generate concerning the principle of due process and judicial autonomy. It concludes that, while AI can support operational tasks under regulated conditions, its intervention should not replace judicial reasoning, the assessment of evidence, or substantive decisions, functions that correspond exclusively to the legal practitioner. It also highlights the need to reduce regional digital divides and enhance training processes in the ethical and responsible use of these technologies, to ensure a digital justice system that is inclusive, efficient, and respects fundamental rights.

Keywords

Artificial intelligence, Judicial Branch, due process, justice, automation

I. Introducción

El vertiginoso avance de las tecnologías digitales y, en particular, de la inteligencia artificial (IA), ha comenzado a impactar de manera significativa diversos sectores de la vida institucional, entre ellos la administración de justicia. En este contexto, se ha vuelto urgente analizar los límites y posibilidades que ofrece la inteligencia artificial generativa en la Rama Judicial, en lo concerniente al

ejercicio de la función jurisdiccional a la luz de principios fundamentales como el debido proceso.

El propósito primordial de la presente investigación radica, mediante una revisión bibliográfica, jurisprudencial y normativa, en examinar el alcance del uso de la inteligencia artificial en la actividad judicial colombiana, especialmente a partir del precedente generado por el fallo del Juzgado Primero Laboral del Circuito de Cartagena, así como de los lineamientos establecidos por el Consejo Superior de la Judicatura mediante el Acuerdo PCSJA24-12243 de 2024. Este análisis busca evaluar los riesgos, beneficios y condiciones bajo las cuales dichas tecnologías pueden integrarse al sistema judicial sin comprometer los pilares esenciales del orden constitucional ni la legitimidad de las decisiones judiciales.

1. Uso judicial de la IA: entre la eficiencia y el debido proceso

El canon 29 de la Constitución Política de Colombia consagra como derecho fundamental el acceso a un juicio en el que se garantice el pleno respeto de las oportunidades procesales para la exposición, contradicción y valoración de las pruebas, con el fin de asegurar una decisión judicial legítima y respetuosa del debido proceso. Al respecto, el artículo 228 *idem* destaca que las normas que rigen el procedimiento, son un instrumento para la adecuada materialización del derecho sustancial.

La delegación parcial de toma de decisiones en sistemas de Inteligencia Artificial ha generado un debate en torno a la legitimidad de las mismas y la forma en la que la ciudadanía pueda aceptar que un sistema carente de rostro, identidad jurídica o atributo humano sea la encargada de conceder o restringir derechos (Ballen, 2023).

Sumado a lo anterior, es menester resaltar que el 30 de enero de 2023, el Juzgado Primero Laboral del Circuito de Cartagena profirió sentencia dentro de una acción de tutela instaurada en favor del menor Salvador Espitia Chávez contra la EPS Salud Total. En la referida providencia, el despacho judicial accedió a las pretensiones del amparo constitucional, disponiendo la exoneración del pago de cuotas moderadoras y copagos a cargo del menor, quien presenta condición de autismo y carece de los medios económicos para asumir tales erogaciones.

Hasta este punto, el fallo podría considerarse una manifestación ordinaria del control de constitucionalidad por vía de tutela, circunscrita a verificar si la decisión se ajusta o se aparta de la línea jurisprudencial de la Corte Constitucional. Sin embargo, adquiere relevancia particular a raíz de un elemento distintivo contenido en la página 5 de la sentencia, específicamente en el tercer párrafo, el cual introduce un aspecto que merece especial atención desde la perspectiva jurídica y judicial, el cual se cita a continuación:

> De otro lado, "atendiendo que la Ley 2213 de 2022 tiene por objeto la incorporación de las TIC en los procesos judiciales", el juez advirtió que haría uso de herramientas de inteligencia artificial generativas (en adelante, IA) para "extender los argumentos de la decisión adoptada".

2. Las preguntas realizadas al aplicativo y las respuestas emitidas por la herramienta ChatGPT 3.5

La sola circunstancia de que el despacho judicial haya acudido a la herramienta ChatGPT como apoyo en la elaboración del fallo, ha suscitado un amplio debate jurídico y académico en torno a la pertinencia, alcances y límites del uso de herramientas basadas en inteligencia artificial (en adelante, IA) en el ámbito del derecho.

Desde ya algunos académicos, tales como José Miguel de la Calle en su artículo «El Juez Artificial» (2024), han expresado su opinión de que, si bien la idea de un juez artificial aún genera temor y enfrenta serios cuestionamientos constitucionales, éticos y técnicos —como la ausencia de conciencia, la imposibilidad de ponderación humana y el riesgo de sesgos algorítmicos—, también reconoce que los jueces humanos no están exentos de sesgos y limitaciones. Por ello, plantea que el uso de inteligencia artificial en la justicia debe abordarse con prudencia, asegurando un uso responsable, gradual y proporcionado, pero sin desconocer su potencial revolucionario para lograr una justicia más accesible, eficiente y oportuna.

El uso de la inteligencia artificial en la administración de justicia puede representar ventajas significativas, como la optimización del tiempo, la automatización de tareas repetitivas y la concentración del talento humano en funciones de mayor valor agregado. En un sistema judicial que aún transita hacia una verdadera digitalización —lejos aún de considerarse tal el simple almacenamiento de expedientes en carpetas digitales de One Drive o el traslado de memoriales por correo electrónico—, la eficiencia se presenta como un argumento legítimo para avanzar en la incorporación de estas herramientas. Sin embargo, ello no puede conducir a justificar su empleo como parámetro para evaluar la calidad de las decisiones judiciales, ni mucho menos asumir, sin fundamento, que la crisis del sistema obedece a la supuesta deficiencia en los fallos humanos. Es claro que los problemas del aparato judicial son complejos y estructurales, derivados de factores como la cobertura, los recursos y las condiciones institucionales, y que los errores judiciales no constituyen la regla general ni explican, por sí solos, las falencias del sistema.

Entonces, ¿debe la IA reemplazar el razonamiento humano? La respuesta es no, la interpretación jurídica, la valoración probatoria y la decisión de fondo deben seguir estando exclusivamente en cabeza del juez, en razón de fundamentos éticos y técnicos, dada la persistencia de errores, imprecisiones y sesgos discriminatorios en los sistemas de inteligencia artificial.

En Colombia, el Consejo Superior de la Judicatura se ha puesto en la tarea de adoptar lineamientos para el uso y aprovechamiento respetuoso, responsable, seguro y ético de la inteligencia artificial en la Rama Judicial, a través del Acuerdo PCSJA24-12243 del 16 de diciembre de 2024, donde, de manera enfática se prohíbe que la IA sustituya al juez en funciones esenciales como la interpretación del derecho, la valoración de la prueba o la adopción de decisiones de fondo, reservadas exclusivamente a la autoridad judicial. No obstante, se autoriza su uso como apoyo en tareas administrativas y operativas, tales como

la organización de expedientes, la asistencia en redacción o la gestión de carga procesal, siempre que el contenido generado por estas herramientas sea revisado críticamente por los funcionarios judiciales.

El Acuerdo también contempla el uso de tecnologías generativas, únicamente como insumo auxiliar y no como fuente autónoma de argumentación jurídica, teniendo como consigna principal, el numeral 1 del artículo 8 de dicha norma: «Evitar el uso de chatbots generales o comerciales de IA en sus versiones gratuitas»; insistiendo en que toda providencia debe responder al razonamiento propio del operador judicial. Finalmente, promueve la capacitación continua del personal judicial en el manejo de estas herramientas, con la Escuela Judicial Rodrigo Lara Bonilla, en articulación y coordinación con la Unidad de Transformación Digital e Informática, para asegurar que su uso se realice con enfoque de derechos y bajo estrictas condiciones de legalidad, ética y responsabilidad institucional.

Esta iniciativa se materializa a través del curso virtual «Inteligencia artificial para la administración de justicia: fundamentos, aplicaciones y buenas prácticas», con una duración total de 50 horas, desarrollado en cooperación con la Universidad de los Andes —a través de su Escuela de Gobierno Alberto Lleras Camargo y la Dirección de Educación Continua—. Para su primer ciclo de formación fueron admitidos 1021 discentes, todos ellos servidores judiciales, quienes al finalizar recibirán una certificación conjunta expedida por la mencionada universidad y la Escuela Judicial Rodrigo Lara Bonilla.

Esto adquiere vital relevancia teniendo en cuenta que el Consejo Superior de la Judicatura adelantó entre el 11 y el 26 de julio de 2024 una encuesta nacional dirigida a los servidores judiciales, donde el objetivo principal de esta iniciativa fue identificar los usos actuales y proyectos en curso que incorporan inteligencia artificial en el interior de la Rama Judicial.

Los hallazgos más relevantes evidencian que el 53,6% de los usuarios de estas herramientas pertenecen a juzgados, destacándose dentro de ellos jueces, profesionales universitarios, secretarios y oficiales mayores. En cuanto al nivel decisorio, el informe señala que

> el 22,1% de los jueces y el 39,9% de los magistrados participantes, quienes representan el nivel decisorio más alto dentro de un despacho judicial, utilizan herramientas de inteligencia artificial, principalmente para consultar información general y jurídica (legislación, jurisprudencia, doctrina), resumir, simplificar y corregir textos, transcribir audios o videos, generar y organizar ideas y, organizar y clasificar datos.

Desde una perspectiva territorial, el distrito judicial de Bogotá se consolida como el principal centro de adopción tecnológica, concentrando el 18,7% de los servidores judiciales que emplean herramientas de IA, seguido por Antioquia-Medellín (12,1%) y Cundinamarca (7,5%), lo que revela una marcada desigualdad regional en la apropiación de estas tecnologías. En contraste, otras zonas como Atlántico-Barranquilla (4,2%), Santander-Bucaramanga (2,9%) y Magdalena-Santa Marta (1,5%) presentan una adopción considerablemente menor.

Esta situación pone en evidencia una brecha digital significativa, razón por la cual el informe concluye que «este panorama resalta la necesidad de fortalecer el acceso a herramientas digitales y la capacitación en el uso de IA, promoviendo una transformación judicial equitativa en todo el país».

II. Cargas del servidor judicial al emplear herramientas de IA en la toma de decisiones

Al respecto la Corte Constitucional se pronunció en la Sentencia T-323 de 2024 donde se revisó el ya mencionado fallo de tutela de segunda instancia del Juzgado Laboral del Circuito, y decidió que no se configuró una vulneración al derecho al debido proceso ni una indebida delegación de la función jurisdiccional, en la medida en que el sistema de inteligencia artificial fue empleado por el juez únicamente con posterioridad a la adopción y fundamentación de la decisión, como herramienta de apoyo y no como sustituto de su criterio judicial.

A través de este precedente jurisprudencial, también le fue ordenado al Consejo Superior de la Judicatura difundir la providencia en todos los despachos judiciales del país, surgiendo así la cartilla informativa «ABC Sentencia T-323 2024», la cual explica la carga tripartita que debe asumir con suma responsabilidad un servidor judicial al momento de usar IA generativa como apoyo en la toma de decisiones, explicada a continuación:

- *Carga de transparencia*: el funcionario judicial que utilice inteligencia artificial en el proceso debe informar expresamente a las partes su uso, justificar su conocimiento y capacitación en la materia, explicar el funcionamiento, capacidades y limitaciones del sistema utilizado, exponer de forma clara y completa la fundamentación relacionada con su empleo, revelar los datos empleados y su incidencia en la decisión, y realizar un análisis de necesidad e idoneidad sobre su aplicación en el caso concreto.
- *Carga de responsabilidad*: El juez o magistrado que utilice herramientas de inteligencia artificial debe estar debidamente capacitado, conocer los riesgos asociados, y justificar el origen, idoneidad y necesidad de su uso. Su responsabilidad recae en verificar que la información generada sea veraz, pertinente al caso, jurídicamente sólida y comprensible. Además, debe asegurarse de que el sistema esté entrenado con datos recientes y adecuados al contexto colombiano, y advertir expresamente cualquier inconsistencia que identifique en la decisión judicial.
- *Carga de privacidad*: es fundamental proteger la reserva de datos personales y sensibles en el sistema judicial, evaluando los riesgos de suministrarlos a herramientas de inteligencia artificial, especialmente si estas son externas o no están autorizadas para funciones judiciales en Colombia, con el fin de prevenir posibles filtraciones (Consejo Superior de la Judicatura (2024c).

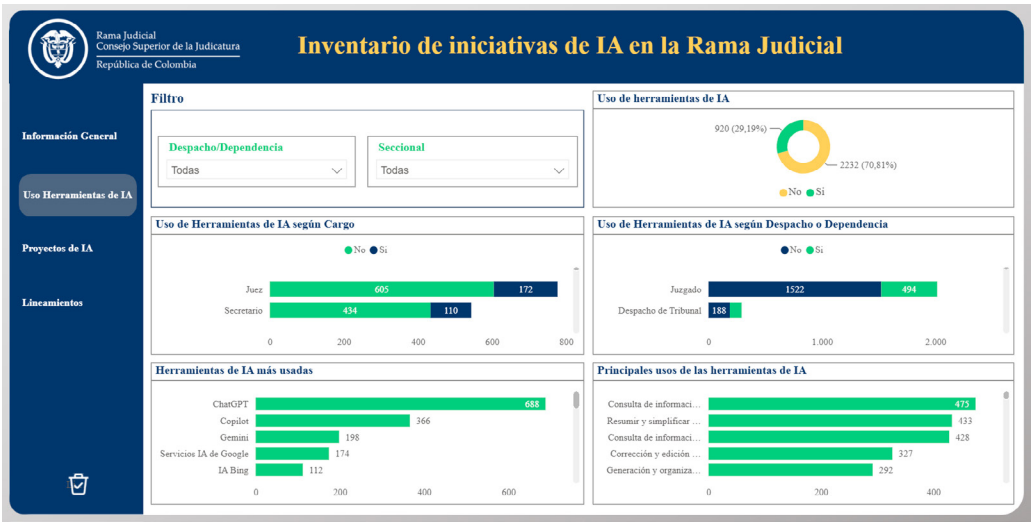
Como consecuencia de lo anterior, se determina que el juez incurre en incumplimiento de sus deberes cuando utiliza inteligencia artificial para realizar funciones propias de su razonamiento judicial; omite informar a las partes sobre su uso, afectando el derecho de contradicción y no verifica rigurosamente

la fiabilidad de la información utilizada, comprometiendo la independencia o imparcialidad por posibles sesgos.

III. Aplicación de la inteligencia artificial generativa en procesos de especialidad civil en el Código General del Proceso

Antes de adoptar una postura frente a las distintas visiones sobre la inteligencia artificial, resulta pertinente abordar la noción de sentencia judicial. Conforme al autor Henry Sanabria Santos (2021, p. 570), las providencias judiciales son los actos procesales del juez, son las decisiones que adopta. Existen dos tipos de providencias: los autos y las sentencias. Las sentencias son aquellas que 1) deciden sobre las pretensiones y excepciones de mérito, en única, primera o segunda instancia; 2) las que deciden incidentes de liquidación, cuando se condenó en abstracto; y 3) las que resuelven los recursos de casación, revisión y anulación de laudos arbitrales.

De acuerdo a los datos de la encuesta referenciada en el acápite anterior, al analizar la distribución por especialidad, se observa que las áreas Civil (19,1 %) y Penal (18,4 %) registraron el mayor número de respuestas, lo cual coincide con la tendencia de mayor participación en los ingresos de procesos judiciales anuales. Esta circunstancia podría estar vinculada con un interés creciente, por parte de los operadores judiciales de la jurisdicción ordinaria —y especialmente de estas ramas—, en explorar el uso de tecnologías de inteligencia artificial (IA) para optimizar su gestión. Teniendo como referente claro está, la tendencia del Distrito Judicial de Villavicencio, donde aún es incipiente el uso de Herramientas IA en las providencias judiciales tal y como se evidencia a continuación:



Fuente: Rama Judicial. (2024). Recuperado de: https://app.powerbi.com/view?r=eyJrIjoiZDMwOWI5MzQtOTJhNC00ZGZmLWJlYjctODdmYmZkNDQ2M2UyIiwidCI6IjYyMmNiYTk4LTgwZjgtNDFmMy04ZGY1LThlYjk5OTAxNTk4YiIsImMiOjR9

De acuerdo a César Augusto Brausín Arévalo (2025), magistrado de la Sala Civil Familia del Tribunal Superior de Distrito Judicial de Villavicencio, las decisiones de plano:

> Se caracterizan por su simplicidad, ya que el legislador define cuándo ocurren y cuál es su contenido, de modo que su motivación es breve: la alusión a la norma que la contempla y la verificación de la ausencia de oposición del demandado, luego de su notificación personal; atributos que facilitan la automatización esta etapa del proceso. (Brausín, 2025, p. 49).

Constituyen ejemplos de decisiones proferidas de plano, en ausencia de oposición por parte del extremo pasivo del proceso, los siguientes casos:

Tabla 1. Autos de plano por silencio del demandado en el CGP

Artículo	Proceso	Contenido
372-2	Rendición provocada de cuentas	Auto que las aprueba y presta mérito ejecutivo
360	Rendición espontánea de cuentas	Auto que las aprueba y presta mérito ejecutivo
409	Divisorio	Auto que ordena división o venta solicitada (*teniendo en cuenta que sea procedente*)
440	Ejecutivo	Auto de proseguir la ejecución
467-4	Adjudicación o realización especial de la garantía real	Auto de adjudicación del bien al acreedor si no se formulan oposición, objeción o petición de remate previo
466-3	Efectividad de la garantía real	Orden de seguir adelante la ejecución

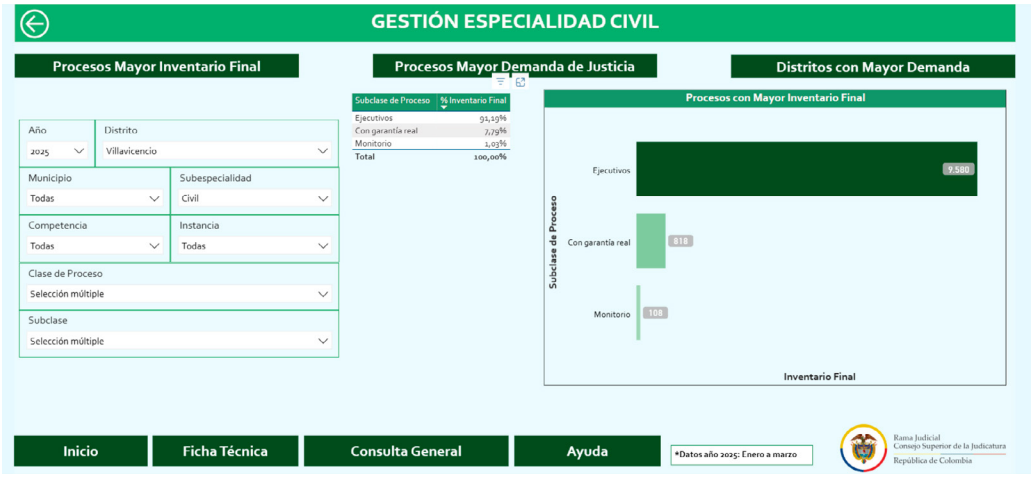
Fuente: Brausín (2025).

Tabla 2. Sentencias de plano por silencio del demandado en el CGP

Artículo	Proceso	Contenido
378-3	Entrega del tradente al adquirente	Sentencia que ordena la entrega
381-2/3	Pago por consignación	Sentencia que declara válido el pago
384-3	Restitución de inmueble arrendado	Sentencia que ordena la restitución
385	Otros procesos de restitución de tenencia	Sentencia que ordena la restitución (*por ejemplo: subarriendo, arriendo de mueble y cualquier otra tenencia diferente, como, por ejemplo,* leasing *y comodato*)
398	Cancelación, reposición y reivindicación de títulos valores	Sentencia que ordena cancelación y reposición
421	Monitorio	Sentencia que condena en el monto solicitado

Fuente: Brausín (2025).

Esta información constituye un insumo valioso para que herramientas de inteligencia artificial generativa asistan en la redacción de sentencias o autos que deban emitirse de plano. De acuerdo con los datos suministrados por la Unidad de Desarrollo y Análisis Estadístico (UDAE), para el período comprendido entre enero y marzo de 2025, el inventario de algunos de los procesos que podrían beneficiarse del uso de IA en la emisión de decisiones de plano se resume en la siguiente figura:



Fuente: UDAE. (2025). https://app.powerbi.com/view?r=eyJrIjoiNTkzM2IxMzgtOTU0N-y00Mjc0LWE3ZTItMTJjMmNhMTg0OTFiIiwidCI6IjYyMmNiYTk4LTgwZjgtNDFmMy04Z-GY1LThlYjk5OTAxNTk4YiIsImMiOjR9

En consecuencia, los datos sugieren que la aplicación de IA generativa en procesos civiles con decisión de plano podría agilizar la emisión de estas providencias, optimizando recursos y beneficiando a los usuarios.

Al respecto, es menester resaltar la opinión de Andrés Villamarín Díaz, juez Segundo Civil del Circuito de Villavicencio, donde cuestiona el uso de esta herramienta no solo en el entendido del uso responsable de la misma, sino:

> En las condiciones necesarias que deben existir por parte del órgano administrativo judicial para que estos avances lleguen a todos los distritos y no sean un motivo más para ensanchar las brechas que, en cuanto a acceso a internet, a equipos idóneos o a tecnologías elementales, hay entre las grandes capitales y las regiones, aquí a diario vemos mensajes sobre fallas de internet o de suministro de energía, otras tantas veces requerimos asistencia con algún aplicativo y quedamos atentos a un «esperemos a ver qué nos dicen de Bogotá, ya se escaló el caso». (Villamarín, 2025, p. 17)

En suma, aunque la inteligencia artificial generativa representa una herramienta prometedora para apoyar la emisión de decisiones de plano en la jurisdicción civil, su implementación debe considerar no solo criterios de eficiencia, sino también condiciones estructurales que garanticen equidad tecnológica entre los distintos distritos judiciales.

IV. Conclusiones

En conclusión, la incorporación de inteligencia artificial generativa en la Rama Judicial colombiana representa un avance significativo hacia la modernización del aparato jurisdiccional, particularmente en tareas operativas como la emisión de decisiones de plano, donde la brevedad argumentativa y la mínima carga interpretativa hacen viable su aplicación como herramienta auxiliar.

No obstante, este proceso debe desarrollarse bajo estrictos límites normativos y éticos que garanticen la preservación del rol esencial del juez en la interpretación del derecho, la valoración probatoria y la adopción de decisiones de fondo, tal como lo establece el Acuerdo PCSJA24-12243 de 2024. Casos como el del Juzgado Primero Laboral del Circuito de Cartagena han evidenciado la urgencia de trazar un marco claro para el uso de estas tecnologías, evitando confusiones sobre su alcance y propósito.

Asimismo, los resultados de la encuesta nacional adelantada por el Consejo Superior de la Judicatura reflejan tanto un creciente interés en el uso de herramientas de IA —con especial énfasis en las jurisdicciones civil y penal— como una preocupante brecha tecnológica entre regiones, que podría profundizar las desigualdades en el acceso efectivo a la justicia. De allí que la capacitación continua de los operadores judiciales, como la ofrecida por la Escuela Judicial Rodrigo Lara Bonilla en alianza con la Universidad de los Andes, resulte indispensable para una apropiación crítica, segura y equitativa de estas herramientas.

Finalizando así con que la eficiencia judicial no puede desvincularse de los principios constitucionales, y menos aún del derecho fundamental al debido proceso, que exige decisiones legítimas, razonadas y adoptadas por autoridad judicial competente. Solo así podrá hablarse de una transformación digital verdaderamente incluyente, garantista y funcional al servicio de la ciudadanía.

Referencias bibliográficas

Ballen, J. (2023). ¿Y qué es eso de ChatGPT? *El Blog del Departamento de Propiedad Intelectual*. https://propintel.uexternado.edu.co/y-que-es-eso-de-chatgpt/

Brausín, C. (2025). La decisión de plano en el Código General del Proceso, como punto de partida para el uso de inteligencia artificial generativa en decisiones judiciales de la especialidad civil. *Amanecer Llanero. Revista Digital del Tribunal Superior del Distrito Judicial de Villavicencio*, (4).

Consejo Superior de la Judicatura. (2024a). *Acuerdo PCSJA24-12243. «Por el cual se adoptan lineamientos para el uso y aprovechamiento respetuoso, responsable, seguro y ético de la inteligencia artificial en la Rama Judicial».*

Consejo Superior de la Judicatura. (2024b). *Inteligencia Artificial en la Justicia Colombiana. Resultados Encuesta Julio 2024. Experiencias de Inteligencia Artificial en la Rama Judicial*. https://www.ramajudicial.gov.co/documents/10635/96912759/Reporte+Ejecutivo+Encuesta+IA.pdf/c5023729-9ad0-ec87-125b-46709ff24533?t=1740177430513

Consejo Superior de la Judicatura. (2024c). *ABC Sentencia T-323 de 2024. JusticIA Lab*. https://escuelajudicial.ramajudicial.gov.co/sites/default/files/ABC_SentenciaIA_T323De2024.pdf

Corte Constitucional de Colombia. (2024). *Sentencia T-323*. Magistrado sustanciador: Juan Carlos Cortés González. https://www.corteconstitucional.gov.co/relatoria/2024/t-323-24

De la Calle, J. (2024). El Juez Artificial. *Legis. Ámbito Jurídico*. https://www.ambitojuridico.com/noticias/columnista-impreso/tic/el-juez-artificial

Sanabria Santos, H. (2021). *Derecho Procesal Civil*. Universidad Externado de Colombia.

Villamarín, A. (2025). Reflexiones sobre el uso de la inteligencia artificial en la Rama Judicial. *Amanecer Llanero. Revista Digital del Tribunal Superior del Distrito Judicial de Villavicencio*, (4).

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 279-292

LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL, ¿SUJETO U OBJETO EN LA RELACIÓN JURÍDICA CIVIL? IMPACTO EN LOS DERECHOS HUMANOS

ARTIFICIAL INTELLIGENCE SYSTEMS: SUBJECT OR OBJECT IN CIVIL LEGAL RELATIONSHIPS? IMPACT ON HUMAN RIGHTS

Vania Jacomino González

Universidad Central «Marta Abreu» de Las Villas

Resumen

Transcurrido el primer cuarto del siglo XXI se ha presenciado por la humanidad un fenómeno que ha repercutido en varios sectores, y el ámbito jurídico no ha sido la excepción, la Inteligencia Artificial está transformado el mundo. El desarrollo que han alcanzado la IA y su propagación masiva en todas las actividades que realiza el hombre, ciertamente ha beneficiado a la sociedad puesto que facilita y agiliza dichas actividades con una mayor eficiencia que el hombre, pero también ha traído consigo la problemática acerca de dónde ubicar a la IA cuando esta interactúa en una relación jurídico civil, así como también la afectación a la que se encuentran sometidos los derechos humanos vulnerados frente al actuar de la IA. A partir de la realización de un estudio bibliográfico que comprendió el análisis de textos clásicos y modernos, publicaciones, artículos, entre otros, se estableció como resultado que la IA van a ubicarse en el elemento objetivo de la relación jurídico civil ya que a pesar de poseer cierta inteligencia y razonamiento sigue siendo un objeto en el tráfico jurídico, por otro lado, se requiere una legislación de carácter nacional que regule a la IA y si interacción con los derechos humanos.

Palabras clave

Inteligencia artificial, objeto, derechos humanos

Abstract

After the first quarter of the 21st century, humanity has witnessed a phenomenon that has impacted various sectors, and the legal field has been no exception: Artificial Intelligence is transforming the world. The development achieved by AI and its widespread propagation across all human activities have certainly benefited society by facilitating and expediting these activities with greater efficiency than humans. However, it has also brought about the issue of where to place AI when it interacts within civil legal relationships, as well as the impact on human rights that are violated by the actions of AI. Based on a bibliographic study that included the analysis of classical and modern texts, publications, articles, among others, it was concluded that AI will be situated within the objective element of civil legal relationships. This is because, despite possessing certain intelligence and reasoning capabilities, it remains an object in legal transactions. On the other hand, national legislation is required to regulate AI and its interaction with human rights.

Keywords

Artificial intelligence, object, human rights

I. Introducción

La relación jurídica ha tenido que transformarse a lo largo del paso de los siglos para adaptarse a esta sociedad que se encuentra en constante cambio, aplicando uno de los principios que rigen en cualquiera de las ramas del Derecho, su carácter dialéctico. La relación jurídica civil se encuentra integrada por tres elementos: sujeto, objeto y causa. El elemento subjetivo entendido como la persona, ya sea individual o ficticia, que va a actuar en la relación jurídica. Por otro lado, el elemento objetivo sería la razón o materia sobre la que se va a desarrollar la relación jurídica, su contenido a tratar. Por último, se encuentra al elemento causal ya que la relación se genera por una causa que la legislación determina.

La sociedad se encuentra en constante transformación y en la rama de la tecnología ha ocurrido toda una revolución en el presente siglo XXI que ha repercutido en varios sectores, y el ámbito jurídico no ha sido la excepción. Los sistemas de inteligencia artificial (IA) han logrado adentrarse en varios ámbitos de la vida de las personas, desde su lugar de trabajo, el desarrollo de los negocios, tareas cotidianas del hogar, entre otros. Y es que la IA facilita y realiza las actividades de forma más eficiente que el hombre, logrando así un alto grado de aprobación y utilización por parte de la población. La contradicción radica a la hora de catalogar a la IA dentro del Derecho Civil puesto que su actuar en la sociedad ha desencadenado la necesidad de determinarla de alguna forma y de atribuirle a su vez las consecuencias de su actuar. Por lo que, el tratamiento que debe aplicarse a la IA se encuentra en estos momentos en análisis y debate entre los teóricos del tema puesto que ya se aprecian diferentes criterios para resolver este conflicto.

Pero este es solo uno de los tantos conflictos que ha traído consigo el uso de la IA de forma masiva en la sociedad, no se pude dejar de lado el impacto con el cual está repercutiendo la IA en lo referente a la vulneración de la cual es objeto los derechos humanos ya sea que deforma intencional se esté utilizando a la inteligencia artificial

Se plantea entonces como objetivos generales:

- Fundamentar teóricamente los presupuestos jurídicos para la determinación de los sistemas de Inteligencia Artificial como elemento de la relación jurídico civil.
- Analizar la repercusión que ha tenido la inteligencia artificial en la violación de derechos humanos viendo a la misma desde dos aristas: objeto y sujeto.

II. Desarrollo

1. La inteligencia artificial

Desde los tiempos antiguos el hombre ha tenido el afán de crear artefactos que tengan semejanzas en el físico o en el comportamiento con el propio ser humano y fue allí donde comenzó la idea de lo que hoy se conoce como inteligencia artificial. Alan Turing, conocido como el padre de la IA propuso una prueba con la finalidad de demostrar la existencia de «inteligencia» en un dispositivo no

biológico, conocida como «test de Turing» se fundamenta en la hipótesis de que, si una máquina se comporta en todos aspectos como inteligente, entonces debe ser inteligente, la misma no comprueba el grado de capacidad de responder a las preguntas correctamente sino de generar respuestas similares a las del ser humano (Ponce Gallegos, 2014).

El apelativo «inteligencia artificial» se debe a McCarthy quien organizó una conferencia en el Darmouth College (McCarthy, Minsky, Rochester y Shannon, 1956) para discutir la posibilidad de construir máquinas «inteligentes»; a esta reunión asistieron científicos de conocida reputación en el área de las ciencias computacionales como: Marvin Minsky, Nathaniel Rochester, Claude Shannon, Herbert Simon y Allen Newell.

1.1 Definición y clasificaciones

La conceptualización de la IA ha sido una labor compleja ya que abarca un contenido extenso y los autores no llegan a un acuerdo para establecer un concepto preciso; teniendo presente que la IA abarca cuestiones de informática, electrónica, programación, mecánica, estadística, cálculo, ingeniería y robótica. McCarthy la definió como «la ciencia e ingenio de hacer máquinas inteligentes, especialmente programas de cómputo inteligentes». La gran mayoría de los conceptos establecen precisamente que la IA es esa máquina con un alto grado de inteligencia que simula de forma bastante precisa el proceso cognitivo humano.

La primera de las clasificaciones es la IA débil que son sistemas diseñados para realizar tareas específicas y limitadas, como el reconocimiento de voz, la identificación de imágenes o la traducción de idiomas; no tienen capacidad de aprendizaje o adaptación por sí mismo y requieren ser programados para realizar una tarea determinada. Su alcance es limitado y no pueden realizar tareas fuera de su campo de especialización. La IA fuerte está diseñada para tener una amplia gama de habilidades cognitivas y capacidad de aprendizaje autónomo. Estos sistemas pueden realizar múltiples tareas y aprenden de forma autónoma a medida que interactúan con el entorno. Tiene la capacidad de razonar, planificar y tomar decisiones complejas en un amplio espectro de situaciones.

2. La personalidad y capacidad en los sistemas de inteligencia artificial

Desde la cuarta revolución industrial, la IA ha logrado alcanzar todos los lugares del planeta, evidentemente no con el mismo alcance y desarrollo, pero si se hace notar su presencia dígase robots en cualquiera de sus generaciones, programas en los Smartphone, en correspondencia con que la IA puede ser tanto un programa como también puede estar presente en el hardware de un sistema. Tal es su alcance que ya se está haciendo referencia en el sector jurídico de otorgarle a la IA personalidad jurídica para que actúe en el tráfico jurídico.

2.1 Posturas con respecto a la regulación jurídica de la inteligencia artificial

La IA no se puede ubicar en ninguno de los parámetros establecidos en el Derecho Civil dígase personas naturales o jurídicas, animales u objetos; entonces la solución a tal cuestión ha derivado a un conflicto entre los teóricos del tema referido a si se puede o no otorgar una personalidad robótica a esos sistemas de IA. Se han establecido tres posibles soluciones para otorgarle la naturaleza legal y protección jurídica a la IA:

- Aquella que estima conveniente conceder a los robots inteligentes una forma de personalidad denominada personalidad electrónica. Esta corriente propone la creación de una nueva categoría jurídica de persona, una suerte *tertium genus* entre persona física y jurídica, denominada persona electrónica, con el objeto de generar un centro de imputación jurídico diferenciado del fabricante, operador, propietario, usuario y demás sujetos vinculados, tratando de crear a largo plazo una personalidad jurídica específica para los robots (Mugas, 2022).
- La que considera que deben recibir la calificación de bien corporal o incorporal. La cualidad de autonomía no resulta innata al robot inteligente, sino que es introducida por sus creadores, de manera que existe un enlace material causal con el comportamiento humano y la actuación del robot. Y si bien es cierto que ellos requieren el dictado de normas específicas, no justifica que reciban un tratamiento distinto a las cosas y demás bienes (Alabart, 2018).
- La que resalta la utilidad de aplicar analógicamente el régimen del esclavo romano con las adaptaciones propias al hecho tecnológico. El vicio congénito de esta opinión es procurar aplicar al supuesto de hecho de los robots un régimen jurídico diseñado para una situación fáctica completamente distinta. Verdaderamente, entre el esclavo y el robot no existe punto en contacto que permita emparentarlos en modo que se pueda utilizar analógicamente su régimen (Alabart, 2018).

2.2 Personalidad electrónica

Con respecto a la primera postura establecida anteriormente que es la que más auge ha tenido en la sociedad, la doctrina actual se ha dividido en dos sectores.

El criterio positivo afirma que es posible determinar en los sistemas de IA la categoría de personalidad ya que esta posee una inteligencia superior a la del ser humano con respecto a la velocidad en la que ejecuta determinados problemas y que su nivel de razonamiento es tal que es posible y necesario que se les otorgue a los sistemas de IA personalidad. Con respecto a lo que plantea el criterio contrario acerca del ADN se resalta que como bien hace mención la historia en tiempos pasados los esclavos aun poseyendo ADN no eran considerados como personas para el Derecho basándose en los argumentos referidos a que por su color de piel eran tratados como cosas, por lo que no eran considerados sujetos a pesar de poseer este requisito por lo que la doctrina no se puede escudar y usar

como argumento este elemento biológico para impedir que la IA adquiera personalidad. Otro fundamento en el que se sustenta este criterio es la existencia de cierto vínculo entre la persona jurídica y la IA puesto que la persona ficticia tiene que existir un fin en el que se sustente su creación y de igual forma sucede con los sistemas de IA la cual se va a encontrar limitada por los fines que le asigne el ordenamiento que autorice su creación. Pero también establecen una diferencia en cuanto a la naturaleza de las personas jurídicas tradicionales y el sujeto «electrónico»: mientras que la primera no es capaz de obrar en el mundo en ausencia de una acción humana, la IA podría llegar a ser potencialmente capaz de «tomar decisiones» a partir de su propio aprendizaje, de forma tal que ellas no puedan ser atribuibles directamente a una acción inmediata de su fabricante o programador (Muñiz, 2018).

La resolución del Parlamento Europeo, del 16/02/2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica, solicita a dicha comisión que proponga definiciones europeas relacionadas con el desarrollo de la robótica en general y la IA para uso civil. Entre las variadas consideraciones, se establece «crear a largo plazo una personalidad jurídica específica para los robots, de forma que como mínimo los robots autónomos más complejos puedan ser considerados personas electrónicas responsables de reparar los daños que puedan causar».

El criterio negativo, niega cualquier posibilidad de que la IA tenga la categoría de personalidad y en consecuencia capacidad puesto que esto es un atributo que exclusivamente les pertenece a las personas, que viene intrínseco con la posibilidad de ser persona y por tanto no se le puede otorgar tal atributo a un ente robótico en el cual no hay presencia de ADN. Los robots como ejemplo de esa IA fuerte entendidos en el campo jurídico como objeto de una relación jurídica, como cosa no puede ocupar la posición de sujeto de la relación jurídica puesto que no tiene los requisitos esenciales para estar en la misma. Y es que, a pesar de la existencia de inteligencia en los sistemas de IA, ésta no posee todos los rasgos que integran a la especie humana dígase la moral, los sentimientos que de cierta forma siempre influyen en el ámbito jurídico y por supuesto que la conciencia que no posee la IA es un elemento de peso a la hora de decidir si otorgarle o no la personalidad.

Con respecto a la responsabilidad que van a adquirir los sistemas de IA se plantean diversas cuestiones problémicas, dígase cómo se podrá determinar si el daño o perjuicio producido por una IA ha sido realizado de forma intencional o culposa si se está en presencia de un dispositivo no biológico en el cual claramente hay inteligencia, pero el mismo carece de conciencia y de voluntad. Por ejemplo, el coche autónomo de Uber que atropelló y mató a una mujer en Tempe, Arizona, el cual circulaba en modo de conducción autónoma, pese a que en el lugar del conductor se encontraba una mujer, a quien se castigó por homicidio negligente pero la causa del accidente lo constituyó una mala programación del algoritmo, el cual, no contemplaba el cruce de personas fuera de los lugares marcados para el cruce peatonal (Fuentes, 2020).

Este es uno de los casos en los que, pensar que la IA es autónoma y puede razonar de la misma manera en que lo harían los humanos, que tomará una

determinación en favor de la vida de las personas, es ingenuo y representa un alto riesgo para la sociedad, y lo mismo puede ocurrir con algoritmos que predicen o toman decisiones con información incompleta o con sesgos en su entrenamiento. (Muñoz, 2024).

Relacionado con ello entra el debate de a quién se le va a exigir dicha responsabilidad, y por ende quién posee la personalidad y capacidad, puesto que se encuentran frente a tres posibles responsables: la persona que creó o fabricó en masa el sistema de IA, ese productor o empresario que de cierta forma va a estar implicado en el suceso; el sujeto que adquirió la IA y le ha dado un determinado uso; o la propia IA, ya que supuestamente se le puede atribuir personalidad y a la vez exigirle responsabilidad.

Otro elemento para tener en cuenta al aplicarle responsabilidad a los sistemas de IA seria la forma de indemnizar o reparar el daño producido a un tercero puesto que con que patrimonio cuenta la IA para responder a esta obligación. Centrémonos por un momento en la Propiedad Intelectual, al adquirir dicha personalidad la IA se consideraría autor o titular en dependencia de la propiedad intelectual creada, pero bajo que fundamentos se aprobaría este actuar cuando es de dominio público que los sistemas de IA ejecutan sus actividades con una base datos incorporadas, entonces como se le puede denominar creatividad a lo que realiza la IA, si está trabajando con elementos ya creados anteriormente por el ser humano. Pero además en donde está la impronta o el sello de autor, donde aparecen el espíritu artístico o lo sentimientos que deja el autor en su obra, una máquina por muy inteligente que sea y a pesar de simular de forma bastante exacta al hombre no tiene forma de dejar tales elementos en una obra.

3. La inteligencia artificial como una creación intelectual

La Inteligencia Artificial, ya sea «débil» o «fuerte», de acuerdo con las clasificaciones que ya se expusieron con anterioridad se relaciona con la propiedad intelectual en tres aspectos principales: la titularidad y protección de los sistemas de Inteligencia Artificial, incluyendo el software, hardware, algoritmos, bases datos o la propia información que los integran, que permiten generar creaciones intelectuales; la titularidad y protección de las creaciones llevadas a cabo por sistemas inteligentes; y la titularidad y protección de los datos y contenidos de los que se nutren los sistemas inteligentes para operar o entrenarse (Vela, 2021).

Si bien España y la Unión Europea tienen regulada la titularidad y protección de los sistemas de IA, esta regulación se realiza de forma singular y no por conjunto, es decir, se encuentra legislado la protección a los algoritmos, *software*, *hardware*, base de datos, y demás información que se le introduzca a la IA, exceptuando evidentemente lo que no se encuentre protegido por el Derecho de Propiedad Intelectual. Este es el patrón que siguen la mayoría de los países otorgándole una protección a los elementos que componen a los sistemas de IA y a un nivel global también se ejecuta de esta manera puesto que estos elementos van a encontrar además una regulación a nivel internacional. Aunque los elementos de los sistemas de IA se encuentren regulados dentro del derecho de propiedad intelectual de esa forma individual y esto de cierta forma ratifica la posición de

que la IA no es más que un objeto que actúa con la sociedad, no es menos cierto que una regulación completa en donde se especifique que la creación intelectual protegida son los sistemas de IA como ese conjunto de software, hardware, base de datos, algoritmo y cualquier información que se introduzca al sistema.

4. Regulación jurídica de la inteligencia artificial

La IA se ha introducido en la sociedad de forma tal que se encuentra inmersa en las actividades que realiza una persona natural pero también en el actuar de las personas jurídicas, dígase en actos sencillamente cotidianos, negocios, en el funcionamiento interno de las empresas, entre otros elementos. Por tanto, no es de asombro que ya el mundo se encuentre regulando todo el actuar de los sistemas de IA y su forma de incidir con respecto a determinadas cuestiones en la sociedad, pues al formar parte de esta como un elemento común es de necesidad su regulación para una mayor seguridad y confiabilidad en el uso de los sistemas de IA. La IA como factor que está actuando de manera activa en diferentes ámbitos de la vida del ser humano no puede moverse en este de forma libre y sin regulaciones, sino que debe tener límites tanto éticos como jurídicos, y el resultado de tal acción no debe ser ponerle al Derecho el papel de freno del desarrollo de los sistemas de IA, sino que este lo que pretende es organizarlo y darle un estatuto de protección a todo lo que rodea a la IA.

Las recientes propuestas normativas realizadas por el Parlamento Europeo en materia de ética, IA y responsabilidad civil de la IA, fruto de los análisis y estudios llevados a cabo durante los últimos años, podrían convertirse o, más bien, son ya en una referencia internacional para su adecuada regulación, cuanto menos desde un enfoque europeo. La UE como guía en ese marco regulatorio del IA de la cual se pondera líder ha establecido el Libro Blanco sobre la IA. Las Resoluciones del Parlamento Europeo de 20 de octubre de 2020, en materia ética y de responsabilidad civil de la IA. Posteriormente, la reciente Propuesta de Reglamento del Parlamento Europeo y del Consejo, de 21 de abril de 2021, estableciendo normas armonizadas sobre IA (Ley de Inteligencia Artificial) y modificando determinados actos legislativos de la Unión, ha apostado por una regulación más profunda y detallada, con claro enfoque a la seguridad y armonización normativa en el seno de la UE (Muñoz, 2024).

Aunque la UE tiene el protagonismo, evidentemente no es el único que ha tomado pasos concretos para una regulación de los sistemas de IA, principalmente teniendo en cuenta que Estados Unidos y China se encuentran en una disputa constante por la supremacía de la IA, aunque los pasos que han dado no solo se centran en la regulación jurídica sino más bien se dirigen al ámbito estratégico y político.

EE. UU. utilizó su *Office of Technology Assesment* para investigar, prever y asesorar al Congreso en el futuro. La Organización Mundial de la Propiedad Intelectual abrió un período de consultas sobre la aplicación de la IA para la generación la propiedad intelectual y abordar el futuro reconocimiento y registro de creaciones generadas por sistemas de IA. Sobre esta materia, el Parlamento Europeo publicó su reciente Resolución de 20 de octubre de 2020, sobre los

derechos de propiedad intelectual para el desarrollo de las tecnologías relativas a la IA.

Japón dispone de sus *Guidelines to Secure the Safe Performance of Next Generation Robots*, así como sus estrategias en robótica mediante la *New Robot Strategy - Japan's Robot Strategy - Vision, Strategy, Action Plan* y la *Headquarters for Japan's Economic Revitalization* de febrero de 2015. China está elaborando un borrador de Reglamento sobre algoritmos de recomendación en Internet que será presentado a consulta pública en breve, en el que se pretende regular un conjunto de obligaciones para los proveedores de recomendación algorítmica. Esta futura norma se unirá a los nuevos marcos jurídicos aprobados en materia de protección de la seguridad de las infraestructuras críticas de información que desarrolla su normativa de ciberseguridad, así como los relativos a la gestión de servicios de información en internet, la seguridad de datos y la protección de la información personal (Vela, 2021).

5. La inteligencia artificial, su impacto en los derechos humanos

Desde el surgimiento de la inteligencia artificial se ha podido observar los aspectos positivos que esta ha traído consigo, pero este factor no deja ocultar los elementos negativos que rodean a la inteligencia artificial y dentro de ello se aprecia el impacto con respecto a los derechos humanos, desde las dos perspectivas que se han desarrollado en el trabajo, es decir de la inteligencia artificial como sujeto y objeto.

A finales de 2018, la Unión de Libertades Civiles de Nueva York, reveló que, en 2017, el *software* de evaluación de clasificación de riesgos (RCA) que utilizaba desde 2013 el Servicio de Inmigración y Control de Aduanas de los Estados Unidos para ayudar a decidir, en procesos de deportación, si un inmigrante debía ser detenido o si podía ser puesto en libertad bajo fianza hasta el momento de la decisión definitiva, había sido manipulado para favorecer las detenciones. El 24 de octubre de 2019, el periódico *El País* publicaba que un equipo de investigadores había demostrado que un algoritmo usado para analizar los riesgos para la salud de millones de pacientes en EE. UU., discriminaba sistemáticamente a la población negra (Asís, 2020).

Estos son solo algunos de los ejemplos que se pueden mencionar del efecto negativo de la IA en su actuar con respecto a la sociedad. De un lado se puede entender que existió un error en el algoritmo o algún otro elemento en la programación y configuración de la IA y entonces el Derecho se encontraría en la disyuntiva ya expresada con anterioridad de a quien responsabilizar entonces por esa violación de los derechos humanos si al programador de la IA, esta persona que la crea, la otra persona que está utilizando la IA y que de cierta forma se está aprovechando de este uso o a la propia IA ya que en este trabajo investigativo se está analizando la posibilidad de otorgarle a la inteligencia artificial una personalidad electrónica para que actúe como sujeto por lo que entonces se estaría en presencia de la violación de los derechos humanos de una persona hacia otra persona y no la de un objeto en el sentido jurídico de estos términos. Pero por otro lado se aprecia también la posibilidad de que los propios seres humanos

y disimiles organizaciones manipulen a la inteligencia artificial para que impacte de la forma que lo deseen en las demás personas existiendo la probabilidad de afectar de manera negativa.

Y es que evidentemente no se aprecia los mismos conflictos cuando es un objeto quien te vulnera los derechos humanos, de acuerdo a errores en su configuración a que ese propio objeto al cual según las estrategias y legislaciones actuales le están otorgando personalidad robótica te vulnere esos derechos humanos y esta situación acarrea nuevas problemáticas a resolver.

Esto demuestra entonces la necesidad de una regulación jurídica no solo en el actuar de la IA y la sociedad, sino que además se tiene que legislar su proceso de creación sobre todo con respecto a los criterios éticos que tienen que regir en una aplicación de IA para que luego la misma no vulnere los derechos legítimos que le corresponden a las personas. Este debería ser el primer factor por lograr para una disminución en los casos de vulneración de derechos humanos, iniciando por el proceso de programación lo que garantizaría que con posterioridad apareciera un número menor de problemáticas.

Por ejemplo, Tay, un bot para Twitter desarrollado por Microsoft (un bot es un programa que ejecuta tareas en internet), capaz de interaccionar automáticamente con usuarios y aprender de ellos. En pocas horas, Tay aprendió expresiones racistas, xenofóbicas, misóginas y fascistas y tuvo que ser bloqueado de la red social. Sin un modelo transparente, es muy difícil hacer valer condiciones sociales con estas tecnologías, como el respeto hacia otros usuarios. Para prevenir que los sistemas hereden tantos prejuicios y que a su vez representen al grueso de la población, se hacen esfuerzos para que exista igualdad de género en los desarrolladores (Hunt, 2016).

Si en el momento de programar a este bot las personas encargadas de tal tarea se hubiesen afiliado a un código de ética para su ejecución y a la vez hubiesen modificados el algoritmo de la IA para que al interactuar con la sociedad solo hubiese escogido los elementos que se acogen a la moral, la buena conducta, el respeto a los derechos humanos, entonces Tay no hubiese vulnerado el derecho humano a la prohibición de discriminación.

El derecho a la reparación integral del daño es un elemento que ya se ha mencionado en varias ocasiones a lo largo del trabajo, pero volvemos a hacer mención del mismo puesto que este se encontraría vulnerado si a la IA se le coloca como sujeto de una relación jurídica civil. Suponiendo esto último expresado en el caso de que un automóvil autónomo atropelle a un peatón, y este por las lesiones recibidas le pida indemnización a esa IA puesto que se le considera persona y por lo tanto tiene la obligación de resarcir el daño causado a los demás, a través de que mecanismo o vía este sistema de IA va a reparar el daño si para el mismo se requiere poseer un patrimonio con el cual hacer frente al mismo. La perspectiva de otorgar personalidad a la IA debe ser seriamente analizada puesto que este aspecto se va a encontrar violando un derecho de las personas. Si nos dirigimos a analizar a la IA como objeto en un proceso de reparación de daños, se presenta entonces otro conflicto, que va a recaer en cuanto a quién repara el daño, la persona que crea el sistema de IA o la persona que haya utilizado la misma.

Como ya ha sido reiterados en varios espacios los sistemas de IA se han generalizado y han alcanzado todos los sectores de la vida del ser humano, siendo su uso extensivo a diferentes áreas laborales, donde llegan a desempeñar funciones en las cuales tienen acceso a información de carácter privado de las personas. La problemática aparece al analizar hasta qué punto se puede confiar en la Inteligencia Artificial para que la misma no divulgue o sencillamente utilice esa información para cuestiones en las cuales el usuario al que pertenece la misma no ha dado su consentimiento para ese actuar. En este caso se estaría vulnerando dos derechos el de acceso a la información ya que es facultad del sujeto a quien y de qué forma puede dar a conocer esa información personal y además el derecho a la protección de esos datos personales puesto que le asiste a la persona el derecho a que donde se encuentren almacenados esos datos sea un lugar confiable y se le protegen de divulgación o utilización no desea por su parte.

6. IA y derechos humanos. Regulación jurídica

En el ámbito de la Unión Europea se han elaborado numerosos documentos en esta línea. Entre todos destaca el Comunicado de 8 de abril de 2019 de la Comisión Europea, Siete requisitos esenciales para lograr una inteligencia artificial fiable. Estos siete requisitos son: 1) la intervención y supervisión humanas (los sistemas de inteligencia artificial deben facilitar sociedades equitativas, apoyando la intervención humana y los derechos fundamentales, y no disminuir, limitar o desorientar la autonomía humana); 2) la robustez y seguridad (los algoritmos deben ser suficientemente seguros, fiables y sólidos para resolver errores o incoherencias durante todas las fases del ciclo de vida útil de los sistemas de inteligencia artificial); 3) privacidad y gestión de datos (los ciudadanos deben tener pleno control sobre sus propios datos, al tiempo que los datos que les conciernen no deben utilizarse para perjudicarles o discriminarles); 4) transparencia (debe garantizarse la trazabilidad de los sistemas de inteligencia artificial); 5) diversidad, no discriminación y equidad (los sistemas de inteligencia artificial deben tener en cuenta el conjunto de capacidades, competencias y necesidades humanas, y garantizar la accesibilidad); 6) bienestar social y medioambiental (los sistemas de inteligencia artificial deben utilizarse para mejorar el cambio social positivo y aumentar la sostenibilidad y la responsabilidad ecológicas); 7) rendición de cuentas (deben implantarse mecanismos que garanticen la responsabilidad y la rendición de cuentas de los sistemas de inteligencia artificial y de sus resultados)[1].

Esta presencia de los derechos humanos en los documentos que se pronuncian sobre la regulación de la inteligencia artificial está presente también fuera de Occidente. Son significativos al respecto, los denominados Principios de Pekín, aprobados en mayo de 2019, por la Academia de Inteligencia Artificial de Pekin (AIAB), la Universidad de Pekin, la Universidad de Tsinghua, el Instituto de Automatización, el Instituto de Tecnología de Computación de la Academia de Ciencias de China, y una liga industrial de inteligencia artificial formada por

1 Disponibles en https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

firmas como Baidu, Alibaba y Tencent. En estos principios se habla del desarrollo sostenible, de la privacidad, de la dignidad, de la libertad, de la autonomía y de los derechos, como referentes para la investigación y el desarrollo de la inteligencia artificial. Así, se afirma que «el desarrollo de la IA debe reflejar la diversidad y la inclusión, y debe diseñarse para beneficiar a la mayor cantidad de personas posible, especialmente a aquellos que, de lo contrario, serían fácilmente descuidados o insuficientemente representado»[2].

La Unesco acordó en julio de 2021 que su Conferencia General debatiera, en su reunión de noviembre, la Recomendación sobre la Ética de la Inteligencia Artificial. En caso de aprobación, será el primer texto de carácter universal sobre la IA basado

> en el derecho internacional y en un enfoque normativo mundial y centrado en la dignidad y los derechos humanos, así como en la igualdad de género, la justicia social y económica, el bienestar físico y mental, la diversidad, la interconexión, la inclusión y la protección del medio ambiente[3].

La regulación jurídica que se va efectuando con respecto a este tema contempla esencialmente algunos aspectos éticos que se deben llevar a cabo con los sistemas de Inteligencia Artificial y los principios que deben regir en la utilización de esta, pero la realidad es que dejan fuera aspectos que también requieren ser regulados, como por ejemplo los ya mencionados con anterioridad en este trabajo.

Otro aspecto relevante en las legislaciones que han sido divulgadas en la actualidad es que la mayoría posee un carácter internacional o en su defecto territorial. Y aunque esto no constituye un error garrafal ni mucho menos si se debe señalar que como la interacción de la IA con los derechos humanos es un elemento que difiere y se transforma de un estado hacia otro, por disimiles factores como puede ser el desarrollo económico, estrategias gubernamentales de los países referentes a la IA, la aceptación de la sociedad, la legislación existente en ese país con respecto a esta materia, entre otros, pues entonces no puede concentrarse entonces la iniciativa legislativa a un alcance global solamente. Cierto es que como estos temas son novedoso y actuales, con pocos años rodeando a la sociedad todavía nos queda muchos errores para corregir y algo de practica para lograr una legislación en materia de IA y derechos humanos acorde a los acontecimientos más recientes en el mundo moderno.

7. Consideraciones generales

La IA ha revolucionado completamente todos los ámbitos de la vida del hombre trayendo consigo consecuencias positivas relativas a la simplificación de tareas que anteriormente se realizaban en un período temporal más extenso y con

2 Disponibles en https://www.baai.ac.cn/blog/beijing-ai-principles

3 El documento propuesto es el Anteproyecto de Recomendación sobre la Ética de la Inteligencia Artificial, elaborado por el Grupo Especial de Expertos (GEE) de la Unesco. Documento Final SHS/BIO/AHEG-AI/2020/4 REV.2. París, 7 de septiembre de 2020.

una mayor eficiencia de la que pueda aportar el ser humano. Pero también, ha presentado diversos aspectos negativos relativos a la creación de los sistemas de IA y la ética que se debe aplicar a la misma; la interacción entre la IA y las personas en materia de Derecho Civil con respecto a cómo se debe tratar el daño causado por un sistema de IA o si en ese caso se le puede catalogar de responsable, más aún si se le puede otorgar personalidad jurídica.

Los sistemas de IA requieren un tratamiento jurídico de objeto, con una regulación especial por constituir un avance tecnológico con un incremento sin precedente, pero es inaudito considerar a la IA como un sujeto de derecho al que se le puede otorgar personalidad y capacidad para que actúe en el tráfico jurídico pues aunque es evidente que esta posee un alto nivel de inteligencia y cierto grado de razonamiento, es incrédulo de nuestra parte otorgarle todos los atributos que por siglos han sido pertenencia exclusiva del hombre. No es posible plantearse la adquisición por parte de la IA de la personalidad electrónica, se estaría deshumanizando al Derecho. Pues por muy semejante que se presente la IA a los seres humanos ya sea en apariencia con los robots humanoides o en el razonamiento y su nivel de inteligencia alcanzado existen aspectos que humanizan al hombre y que la IA no posee hasta el momento, nos referimos a cualidades espirituales, sentimentales, morales, de identidad, afectividad.

Por otro lado, la interacción presente entre los derechos humanos y la Inteligencia Artificial es un tema preocupante en la actualidad debido a la dificultad que tare consigo que un sistema de IA vulnere los derechos de un individuo afectando su esfera individual y la problematiza se presenta a la hora de resolver sobre quien recae el deber de responder por tal actuar. Esto se agrava al no existir una regulación jurídica en cada país, sino que la generalidad es acogerse a las normas internacionales.

III. Conclusiones

Primera. Teniendo en cuenta el desarrollo que ha alcanzado la IA y con la rapidez que se ha extendido por todas las actividades del hombre, la gran mayoría de los países se han dedicado a establecer regulaciones o a designar estrategias para organizar y darle un respaldo legal a todo el actuar que conlleve su presencia, con el fin de brindar una mayor seguridad y protección en el ámbito en que se esté desenvolviendo.

Segunda. Con respecto a los criterios que se han establecido alrededor de la determinación de la IA dentro de una relación jurídica civil predominan dos, por un lado, se encuentran los que plantean la posibilidad de otorgarle la denominada personalidad robótica para que puedan interactuar en la relación jurídica como un sujeto de derecho; por otro lado, están los que se oponen rotundamente a otorgarle esa personalidad a la IA puesto que se está en presencia de una máquina que por muy inteligente que sea y a la vez que pueda llegar a simular el razonamiento del hombre, no puede poseer ni conciencia ni los sentimientos que están implicados en el actuar del ser humano a diario, acotando que no posee un patrimonio con el cual enfrentar su actuar.

Tercera. Los sistemas de IA al encontrarse inmersos en una relación jurídica civil se establecen como el elemento objetivo de la misma, siendo la razón o materia sobre la que va a tratar. Conservando la posición del elemento subjetivo a las personas, naturales o jurídicas, puesto que en ambas intervienen los seres humanos, y es que el Derecho desde la antigüedad y debería permanecer de esa forma, es una ciencia que gira alrededor de la figura del ser humano y en función de alcanzar la mayor protección y seguridad para este.

Cuarta. La interacción entre la IA y los derechos humanos es motivo de preocupación en la sociedad, puesto que no existen los mecanismos apropiados para resolver tal conflicto y esto se agrega a la situación de la legislación en esta materia, lo cual da a entender la necesidad presente de un análisis más detallado y de una regulación acorde a la realidad que vive la sociedad.

Referencias bibliográficas

Alabart, S. D. (2018). *Robots y responsabilidad civil*. Reus.

Asís, R. D. (2020). Inteligencia Artificial y Derechos Humanos. *Materiales de Filosofía del Derecho*.

Fuentes, V. (2020). Acusada de homicidio la conductora de Uber cuyo coche autónomo atropelló a una mujer en Arizona. *Motorpasión*. https://www.motorpasion.com/seguridad/acusada-homicidio-conductora-coche-autonomo-uber-que-causo-atropello-mortal-arizona

Grigore, A. E. (2022). Derechos humanos e Inteligencia Artificial. *Ius et Scientia*, 164-175.

Hunt, E. (2016). *Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter*.

McCarthy, J., Minsky, M., Rochester, N. y Shannon, C. (1956). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*.

Mugas, P. E. (2022). El robot inteligente y su categorización jurídica. *Revista Blockchain Inteligencia Artificial*, (4), 29-46.

Muñiz, C. (2018). Para nosotros, para nuestra posteridad, y para todos los robots del mundo que quieran habitar el suelo argentino. ¿Puede la inteligencia artificial ser sujeto de derecho? *Revista Código Civil y Comercial*, (2), 1-7.

Muñoz, F. S. (2024). *Inteligencia Artificial: Una mirada crítica al concepto de personalidad electrónica de los robots*. Universidad de Guanajuato.

Ponce Gallegos, J. C. (2014). *Inteligencia Artificial*.

Quirós, J. J. (2022). Derechos Humanos e Inteligencia Artificial. *Revista Dikaiosyne*, (37), 140-163.

Vela, J. M. (2021). *Derecho de la Inteligencia Artificial. Un enfoque global de responsabilidad desde la ética, la seguridad y las nuevas propuestas reguladoras europeas*. (Tesis doctoral). Universidad de Valencia.

INFORMÁTICA Y DERECHO

REVISTA IBEROAMERICANA DE DERECHO INFORMÁTICO
(SEGUNDA ÉPOCA)

FEDERACIÓN IBEROAMERICANA DE ASOCIACIONES
DE DERECHO E INFORMÁTICA

ISSN 2530-4496 – N.º 18, 2026, PP. 293-305

INTELIGENCIA ARTIFICIAL Y ÉTICA: DESAFÍOS PARA LOS DERECHOS HUMANOS

ARTIFICIAL INTELLIGENCE AND ETHICS: CHALLENGES FOR HUMAN RIGHTS

Bárbara Rosa Chinea Cárdenas

Universidad Central «Marta Abreu» de Las Villas

Resumen

La investigación se focaliza en el análisis de los desafíos éticos que plantea la IA en relación con la protección de los derechos humanos fundamentales. El objetivo principal es identificar medidas y propuestas para integrar la IA en la sociedad sin que ello implique afectaciones negativas en derechos, como la equidad, no discriminación y la protección de datos personales. Para ello se realizará una caracterización de los principales desafíos éticos que enfrenta la IA y a la vez se van a proponer medidas concretas para lograr la reducción de la discriminación algorítmica, el fortalecimiento de la protección de datos y el mejoramiento de la transparencia y la rendición de cuentas en sistemas inteligentes. La investigación emplea un enfoque cualitativo, basado en el análisis bibliográfico de fuentes relevantes y marcos regulatorios internacionales. Este proyecto muestra gran relevancia, principalmente porque responde a la necesidad urgente de armonizar el avance tecnológico con la protección efectiva de los derechos humanos. Los resultados de dicha investigación buscan guiar a desarrolladores, legisladores y a la sociedad civil en general hacia una adopción responsable y ética de la inteligencia artificial, asegurando con ello la innovación sin menoscabo de la dignidad y los derechos de las personas.

Palabras clave

Inteligencia artificial, derechos humanos, ética

Abstract

The research focuses on the analysis of the ethical challenges posed by AI in relation to the protection of fundamental human rights. The main objective is to identify measures and proposals to integrate AI into society without causing negative impacts on rights, such as equity, non-discrimination, and the protection of personal data. To this end, a characterization of the main ethical challenges faced by AI will be carried out, while concrete measures will be proposed to reduce algorithmic discrimination, strengthen data protection, and improve transparency and accountability in intelligent systems. The research employs a qualitative approach, based on the bibliographic analysis of relevant sources and international regulatory frameworks. This project is highly relevant, primarily because it addresses the urgent need to harmonize technological advancement with the effective protection of human rights. The results of this research aim to guide developers, policymakers, and civil society in general toward a responsible and ethical adoption of artificial intelligence, thereby ensuring innovation without undermining human dignity and rights.

Keywords

Artificial intelligence, human rights, ethics

I. Introducción

La Inteligencia Artificial (IA), con su rápido desarrollo ha logrado un cambio sustancial en la sociedad, pues ha logrado integrarse prácticamente en todos los ámbitos, como son el de la salud, la educación, la cultura, y por supuesto que el Derecho no podía estar exento a estas transformaciones. Si bien la IA desde un primer momento ha ofrecido innumerables beneficios al ser humano, también ha presentado significativos desafíos, siendo de los más polémicos los referidos a los derechos humanos, pues son considerados el marco fundamental para evaluar el impacto que ha tenido la IA en el día a día del hombre, debido a su capacidad inherente para afectar, de manera directa y profunda, los derechos y libertades fundamentales de las personas.

Cuando se habla de IA y derechos humanos, hay un enfoque especial a los aspectos que tratan la diversidad, la ética y la privacidad, pues se puede decir que son los más afectados en la vida cotidiana, a través de las decisiones tomadas por algoritmos que pueden discriminar a grupos determinados o incluso a través de la recogida masiva de datos personales.

A partir de ello, se deriva el conflicto siguiente: La equidad, la no discriminación y la protección de datos personales, constituyen derechos fundamentales inherentes a cada persona física. Sin embargo, para el desarrollo y la aplicación de la inteligencia artificial en la sociedad, estos son violados en ocasiones varias, causando con ello, perjuicios graves a los individuos.

Teniendo lo anterior como base, surge la interrogante, ¿Cómo se puede lograr una mayor protección de la ética en el desarrollo y utilización de la IA?

Dicho artículo se desarrolla con el objetivo determinar medidas que ayuden a implementar la IA en la sociedad sin causar afectaciones en los derechos fundamentales del hombre. Utilizando para ello un enfoque cualitativo y métodos como el análisis bibliográfico.

La estructura constará con dos capítulos, refiriéndose en un primer momento a los fundamentos conceptuales de dos figuras esenciales: la IA y los Derechos Humanos, con énfasis en la dimensión ética de estos. También se abordará la relación existente entre la IA y la ética, los marcos regulatorios respecto al tema y la necesidad inmediata de disminuir las violaciones existentes a los derechos. En un segundo momento, se explicarán los principales desafíos existentes, las afectaciones causadas por estos, así como, las medidas que se consideran necesarias para eliminar, o al menos, lograr disminuir dichas dificultades.

II. Desarrollo

1. Fundamentos conceptuales

1.1 Conceptualización de la IA

La inteligencia artificial es una disciplina científica y tecnológica enfocada en el desarrollo de sistemas capaces de realizar tareas que, en condiciones normales, requieren inteligencia humana, tales como el aprendizaje, el razonamiento,

la percepción, la toma de decisiones, el reconocimiento de patrones y el procesamiento del lenguaje (Russell y Norvig, 2021, p. 2). Estos sistemas pueden ser diseñados para funciones específicas, conocidos como IA estrecha por ejemplo asistentes virtuales, o aspirar a una IA general, que igualaría la versatilidad cognitiva humana, aunque esta última aún no se ha logrado (McCarthy, 2007). La IA opera mediante algoritmos que procesan grandes volúmenes de datos, empleando técnicas como el aprendizaje automático y las redes neuronales, lo que permite a ciertos sistemas mejorar autónomamente su desempeño mediante la experiencia Goodfellow, Bengio y Courville 2016. Según Stuart Russell y Peter Norvig, la IA estudia agentes inteligentes que perciben su entorno y actúan para maximizar sus probabilidades de éxito (Russell y Norvig, 2021). Por su parte John McCarthy, pionero del término en 1956, la definió como «la ciencia e ingeniería de hacer máquinas inteligentes». Instituciones como la IEEE y la Unión Europea destacan que estos sistemas exhiben comportamientos inteligentes al interactuar con su entorno de manera autónoma (Institute of Electrical and Electronics Engineers [IEEE], 2023). En esencia, la IA combina enfoques técnicos y cognitivos, consolidándose como un campo interdisciplinario con aplicaciones en casi todos los ámbitos de la sociedad moderna.

1.2 Concepto de derechos humanos: ética

Son entendidos como garantías fundamentales inherentes a todas las personas, que buscan proteger su dignidad y promover su bienestar y desarrollo pleno (Fernández, 1984). No son concesiones de los Estados sino el resultado de un proceso histórico de luchas y demandas sociales por la dignidad humana (Comisión Presidencial Coordinadora de la Política del Ejecutivo en Materia de Derechos Humanos [Copredeh], 2011). No fue hasta la «Declaración Universal de los Derechos Humanos» de 1948, que se marcó el inicio de la institucionalización de los derechos humanos convirtiéndolos de esta forma en parte del Derecho Internacional Público y elevándolos a obligaciones internacionales para los Estados (Copredeh, 2011).

Los derechos humanos, además, pueden verse desde varias dimensiones, en el caso de la dimensión jurídica, al constituir los derechos humanos como normas, estos son reclamables frente al Estado, proporcionando así los medios para su eficaz y posible protección, mientras que, según la dimensión ética se expone que los derechos humanos tienen una profunda raíz moral y son considerados derechos morales (Beuchot Puente, s.f.), se basan en valores y principios éticos como lo son la dignidad, la libertad, la igualdad, la autonomía individual y la justicia (Fernández, 1984).

Por su parte, la ética, en un sentido amplio, se entiende como un conjunto de valores y normas morales y jurídicas (Fernández, 1984) que actúan como criterios de decisión fundamentales que guían la conducta humana y profesional. Su propósito es asegurar que las diversas disciplinas y tecnologías estén al servicio de todos los seres humanos, respetando a su vez los derechos humanos de todos los grupos involucrados.

1.3 Relación de la IA y la ética

En el contexto de la inteligencia artificial, los derechos humanos son verdaderamente relevantes por lo que es necesario comprender como la tecnología interactúa con ellos, sobre todo en temas de privacidad, transparencia, equidad y no discriminación. Por ello, entre los principios éticos claves, especialmente en aspectos relacionados con la inteligencia artificial, se encuentran: la justicia y la equidad, que requiere que se eviten sesgos injustos y que se promueva un trato equitativo para todos, sin discriminación (Agencia Española de Protección de Datos [AEPD], 2020); la autonomía, respaldando así, el derecho de las personas a tomar decisiones sobre sus datos (Tecnología Educativa, 2024); la responsabilidad y rendición de cuentas, que establece la necesidad de mecanismos claros para asegurar la responsabilidad por los errores o daños causados por sistemas tecnológicos (Grupo de expertos de alto nivel sobre la IA, 2019); y la privacidad y gestión de datos, cuya base consiste en el respeto a la privacidad de cada individuo, garantizando de esta forma la calidad, integridad, así como, el acceso legítimo a la información (Guaña-Moya y Chipuxi-Fajardo, 2023).

Es necesario entender que la irrupción de la inteligencia artificial ha convertido la ética en un área de preocupación fundamental. Se considera que los dilemas éticos comienzan cuando a través de la IA se reflejan o amplifican sesgos presentes en los datos que le fueron introducidos en un primer momento, ocasionando discriminación, violaciones a la privacidad de las personas e incluso afectaciones a la dignidad humana (Wolk Soft, 2025).

Ante dichos principios claves, necesarios para la correcta utilización de aplicaciones de inteligencia artificial, se ha vuelto de suma importancia la adaptabilidad o incluso la creación de marcos regulatorios para una IA fiable.

1.4 Marcos regulatorios

Entre las principales normativas para lograr la correcta regulación de la IA, respetando los derechos humanos de cada individuo, se encuentra el Reglamento General de Protección de Datos (RGPD) de la Unión Europea, el cual establece artículos relativos a la protección de las personas físicas en cuanto a sus datos personales y su libre circulación (AEPD, 2020), teniendo en cuenta principios como la licitud, lealtad, transparencia y exactitud. De igual forma, se aprecian leyes nacionales precisamente sobre la protección de datos personales, como lo es la Ley 1581/2012 de Colombia o la Ley 149/2022 de Cuba, que, si bien no fueron redactadas teniendo en cuenta en su totalidad el desarrollo actual de la IA, buscan garantizar también los principios de protección de datos, aplicables precisamente al desarrollo y uso de la propia inteligencia artificial (Morales Cáceres, 2020).

Otra importante normativa es la Declaración Universal de los Derechos Humanos (DUDH), la cual, si bien no es una regulación específica para la inteligencia artificial, sus valores de libertad, igualdad y no discriminación (Copredeh, 2011) son la base ética para cualquier desarrollo tecnológico.

Finalmente, se encuentra una de las normativas jurídicas más conocidas a nivel mundial, la legislación de la Unión Europea, considerada la primera ley

integral sobre IA del mundo. Su prioridad es garantizar que los sistemas de IA sean seguros, transparentes, trazables, no discriminatorios y respetuosos con el medio ambiente. Siguiendo dicho objetivo, aseguran que los sistemas de inteligencia artificial serán supervisados por personas, en lugar de la automatización, para evitar de esta forma resultados perjudiciales (Parlamento Europeo, 2025).

El desarrollo de la IA se encuentra en una fase inicial, lo que lo convierte en un momento clave para asegurar que estos avances tecnológicos sean éticos, brinden protección a los datospersonales, privacidad, equidad y no discriminación. Es de vital importancia equilibrar dichas tecnologías, en especial la IA, con el respeto a los derechos fundamentales del hombre, logrando un desarrollo que no suponga un desafío eventualmente para la dignidad humana (Morales Cáceres, 2020). Por tanto, la reflexión ética, así como, las propuestas de disposiciones legales acordes al tema, deben ser continuas para lograr que la sociedad se adapte correctamente y sin daños a los cambios tecnológicos que cada vez son mayores y más abarcadores.

2. La IA desde el ámbito de la ética

2.1 Principales desafíos éticos en relación a la IA

Como bien se ha planteado, con el creciente desarrollo de la Inteligencia Artificial, los derechos humanos, sobre todo, aquellos relacionados con la ética se han vuelto un tema que requiere atención y reflexión debido a los diversos y complejos desafíos que presenta.

2.1.1 Sesgos y discriminación

Uno de los problemas más conocidos es el sesgo algorítmico, donde los algoritmos, a pesar de ser fórmulas matemáticas en principio neutrales y objetivas, pueden reproducir e incluso amplificar prejuicios humanos de género, raza o edad (Equipo Laboratoria, 2023), causando con sus decisiones afectaciones a los derechos fundamentales de las personas.

Los sesgos pueden ser estadísticos (errores en la recopilación de datos), culturales (derivados de la sociedad y el lenguaje) o cognitivos (basados en creencias individuales) (Morales Cáceres, 2020). Como la IA aprende de grandes volúmenes de información; si estos datos son parciales o limitados en diversidad, los resultados de la IA también lo serán, pues estos sesgos surgen principalmente cuando los datos utilizados para entrenar los algoritmos reflejan patrones históricos de discriminación o son desequilibrados y poco representativos (Equipo Laboratoria, 2023). Sin embargo, también pueden surgir de la información obtenida a partir de la interacción con sus usuarios, en especial en el famoso *machine learning* Esto es fácilmente demostrable a través de los sistemas de reconocimiento facial, los cuales presentan dificultades para identificar personas con tonos de piel más oscuros o de raza negra o asiática, con tasas de error más altas en comparación con personas de raza blanca (Naranjo Faccini, 2024). Un ejemplo de ello fue cuando el *software* de reconocimiento facial de Google Photos etiquetó a una mujer afroamericana como «gorila» (Morales Cáceres, 2020).

2.1.2 Privacidad y gestión de datos

El desarrollo de inteligencia artificial utiliza como uno de sus métodos principales, el análisis de grandes volúmenes de datos personales, lo cual puede traer consigo la vulneración del consentimiento informado e incluso en muchas ocasiones la no protección de datos personales.

Esto ocurre principalmente debido a que la IA está estrechamente relacionada con la recopilación y uso de datos sensibles, ya que a menudo requiere acceso a categorías especiales de datos (origen étnico o racial, opiniones políticas, convicciones religiosas, datos de salud) (AEPD, 2020). En cuanto, al uso de reconocimiento facial o de huellas dactilares también surgen preocupaciones porque en varias ocasiones lleva a la vigilancia masiva y a la recolección de datos sensibles con impactos negativos en comunidades minoritarias y grupos vulnerables (Equipo Laboratoria, 2023).

Por su parte, las IA generativas pueden autoentrenarse con las consultas y respuestas de los usuarios, lo que implica el almacenamiento y uso de información potencialmente personal o confidencial, con el riesgo de que se filtren datos sensibles o se utilicen sin consentimiento explícito para generar respuestas a otros usuarios (Naranjo Faccini, 2024).

2.1.3 Refuerzo de estereotipos y desigualdades sociales

Este desafío está intrínsecamente ligado al sesgo algorítmico, pero se enfoca en cómo los algoritmos de IA pueden amplificar las desigualdades y los prejuicios ya existentes en la sociedad, sí son entrenados con datos sesgados que reflejan prejuicios históricos o discriminación (Wolk Soft, 2025), afectando así a grupos vulnerables y perpetuando la discriminación sistémica.

Una manifestación clara de lo planteado puede ser en la marginación de grupos vulnerables (Equipo Laboratoria, 2023), pues los algoritmos entrenados con datos sesgados que reflejen prejuicios históricos, pueden llevar a la marginación de grupos específicos, como minorías, personas con discapacidad, ancianos o migrantes. Otra manifestación es la reproducción de los estereotipos ya que la inteligencia artificial puede reforzar estereotipos de género y raza en áreas como la publicidad (Equipo Laboratoria, 2023), e incluso en plataformas de citas en línea (criticadas por sesgos raciales en la elección de parejas).

2.1.4 Neutralidad tecnológica y acceso equitativo

Este desafío emerge en el contexto de la interacción entre la tecnología y las políticas de acceso, particularmente en servicios fundamentales como las telecomunicaciones, donde la IA puede influir en la forma en que los usuarios acceden y utilizan la información.

En principio se plantea que la neutralidad tecnológica es aquella que busca garantizar que todos los usuarios tengan acceso equitativo a diferentes plataformas y aplicaciones, sin preferencias comerciales o técnicas impuestas por los proveedores de servicios. El dilema, en sí, surge cuando los planes de telefonía móvil incluyen acceso gratuito a plataformas populares como lo son Facebook y WhatsApp, lo cual, si bien es cierto que beneficia la asequibilidad

para poblaciones de bajos ingresos, a la vez vulnera el principio de neutralidad tecnológica, pues se impone un sesgo y favorece determinadas marcas o aplicaciones (Naranjo Faccini, 2024).

A partir de ello se aprecia claramente la necesidad existente en la actualidad de establecer una serie de medidas para eliminar o al menos disminuir, las violaciones que se ocasionan a los derechos fundamentales de las personas con el desarrollo y uso de la IA.

2.2 Medidas a tomar

El objetivo, no es eliminar el uso de la inteligencia artificial, pues es evidente en la cotidianeidad los beneficios que le otorga al hombre, sin embargo, no podemos dejar que nos ciegue y permitir la violación de derechos fundamentales inherentes a las personas, sino que hay que estudiar en profundidad y colaborar entre todos los especialistas para lograr que la IA y el respeto por la ética y los derechos humanos en general puedan coexistir entre sí. Es por ello que teniendo en cuenta la relación existente entre la ética y la IA explicada con anterioridad, así como, los marcos regulatorios que abordan el tema y los desafíos identificados en la doctrina, se realiza una propuesta de medidas a considerar para mitigar los daños causados a la sociedad.

2.2.1 Abordar sesgos y promover la no discriminación y la equidad

Para abordar los sesgos y promover la no discriminación y la equidad se propone:

2.2.2 La creación de conjuntos de datos inclusivos y diversos

Es fundamental que los algoritmos de IA se entrenen con datos que representen la diversidad de género, etnia y edades de la población, y que estos conjuntos de datos se actualicen continuamente para reflejar los cambios demográficos y sociales, apoyando de esta manera a la reducción de prejuicios y la discriminación.

2.2.3 Que se conformen equipos de desarrollo diversos

Involucrar a profesionales con diferentes antecedentes, géneros, etnias y experiencias en los equipos que desarrollan la IA, logrando así un equipo diverso, el cual tendrá opiniones más variadas y una mejor preparación para identificar potenciales fuentes de sesgo inconsciente y desarrollar soluciones más equitativas e inclusivas.

2.2.4 Elaborar una estrategia para Protección de grupos vulnerables

Las decisiones éticas en IA deben garantizar el acceso igualitario a servicios y oportunidades, especialmente para grupos vulnerables (personas con discapacidad, ancianos, migrantes, minorías) promoviendo de esta forma su dignidad e inclusión.

2.3 Fortalecer la protección de datos personales

En aras de fortalecer la protección de datos personales será preciso:

2.3.1 El cumplimiento de principios de protección de datos

Asegurar que el tratamiento de datos personales en IA se realice conforme a principios como lo son la legalidad, consentimiento, finalidad, proporcionalidad, calidad y seguridad.

2.3.2 La minimización de datos

Los datos personales utilizados deben ser adecuados, pertinentes y limitados a lo necesario para cada fin. Esto incluye depurar y suprimir datos de entrenamiento cuando ya no sean necesarios o no sean estrictamente relevantes.

2.3.3 La gestión del riesgo

Realizar un análisis continuo del riesgo que puede traer consigo el uso de una IA determinada para los derechos y libertades de las personas. Las Evaluaciones de Impacto en Protección de Datos Personales (EIPD o DPIA) son herramientas preventivas esenciales que deben realizarse antes del tratamiento de datos que entrañen un alto riesgo, para identificar, evaluar y gestionar dichos riesgos.

2.3.4 La privacidad desde el diseño

Incorporar medidas técnicas y organizativas para la protección de datos desde las primeras fases del desarrollo de la tecnología, asegurando que la privacidad sea un componente esencial y no un añadido posterior.

2.4 Mejorar la transparencia y explicabilidad de la IA

A fin de mejorar la transparencia y explicabilidad de la IA deben lograrse:

2.4.1 Modelos de IA transparentes y explicables

Desarrollar sistemas que permitan identificar por qué se toman ciertas decisiones y cómo se llegó a un resultado específico, incluso si no se requiere «abrir» completamente la «caja negra» del algoritmo. El objetivo es que las personas entiendan el proceso y puedan confiar en el sistema.

2.4.2 Información clara y concisa

Los responsables deben informar a los usuarios de manera detallada, sencilla, expresa e inequívoca sobre la finalidad del tratamiento de sus datos, sus destinatarios, el carácter obligatorio o facultativo de la información, y la posibilidad de ejercer sus derechos. En el caso de decisiones automatizadas, se debe informar explícitamente de esta circunstancia y del derecho a oponerse o solicitar intervención humana.

2.4.3 Certificación y auditorías externas

Implementar mecanismos de certificación por terceros independientes para acreditar la calidad y confiabilidad de las soluciones de IA y el cumplimiento del Reglamento General de Protección de Datos (RGPD). El monitoreo y la auditoría regular de los algoritmos son fundamentales para identificar y corregir cualquier sesgo que pueda surgir a lo largo del tiempo.

2.5 Asegurar la rendición de cuentas y la responsabilidad

Para asegurar la rendición de cuentas y la responsabilidad se necesita:

2.5.1 Supervisión humana

Aunque la IA puede optimizar procesos, la supervisión humana sigue siendo crucial en las decisiones finales para garantizar que sean justas y éticas. Se debe permitir que un operador humano pueda ignorar el algoritmo si es necesario y se deben documentar estas incidencias.

2.5.2 Mecanismos de responsabilidad claros

Establecer quién es responsable cuando un sistema de IA comete un error o causa un daño, ya sea el desarrollador, la empresa que implementó la IA o la propia máquina. Esto requiere definir los límites de responsabilidad de cada actor en el ecosistema de la IA.

2.5.3 Delegado de Protección de Datos (DPD)

El nombramiento de un DPD, incluso cuando no es obligatorio, puede ser de gran utilidad para gestionar el riesgo y aplicar eficazmente los mecanismos de responsabilidad proactiva y transparencia.

2.6 Marcos éticos y regulatorios robustos

Para fortalecer los marcos éticos y regulatorios se precisa:

2.6.1 Desarrollo de normativas estrictas

Los gobiernos deben establecer marcos regulatorios más estrictos para guiar el desarrollo y uso de la IA, como la Ley de Inteligencia Artificial de la Unión Europea, que busca garantizar que las tecnologías sean seguras, transparentes y no discriminatorias.

2.6.2 Colaboración multidisciplinar

La resolución de los dilemas éticos de la IA requiere un enfoque ético multidisciplinario que involucre a desarrolladores, legisladores, académicos y expertos en derechos humanos. La reflexión ética constante es fundamental para abordar los desafíos planteados por la IA en diferentes contextos.

III. Conclusiones

Para concluir puede afirmarse que, en primer lugar, la Inteligencia Artificial, es considerada un sistema tecnológico autónomo y adaptativo, capaz de realizar tareas para las que anteriormente era necesaria la inteligencia humana. Además, los derechos humanos tienen una profunda raíz moral y son considerados derechos morales, mientras que la ética en un sentido amplio, se entiende como un conjunto de valores y normas morales y jurídicas (Fernández, 1984) que actúan como criterios de decisión fundamentales que guían la conducta humana y profesional.

La IA, a pesar de sus beneficios transformadores, genera grandes conflictos éticos pues impacta de forma directa sobre derechos humanos fundamentales. Debido a ello la ética se ha convertido actualmente en un área de preocupación fundamental. Se considera que los dilemas éticos comienzan cuando a través de la IA daña precisamente derechos como la privacidad, la no discriminación y la equidad de las personas afectando no solo a la sociedad en general sino también en muchas ocasiones la dignidad humana.

En segundo lugar, la regulación existente actualmente, ayuda en algunos de los desafíos planteados, por lo que se puede considerar un punto de partida, sin embargo, aún es insuficiente. Para lograr la completa protección de los derechos humanos de las personas es necesaria una adaptación normativa de forma continua, que vaya evolucionando paso a paso con la tecnología a pesar de su dificultad, pues precisamente de eso se trata el carácter dinámico del Derecho, este tiene que responder a las necesidades de la sociedad. Además, se hará necesaria la colaboración entre especialistas de múltiples disciplinas logrando con ello un equilibrio entre la innovación y el propio Derecho.

Así, entre los desafíos éticos más importantes se encuentran los sesgos algorítmicos, la gestión de datos personales, así como el refuerzo de desigualdades sistémicas y la neutralidad tecnológica, trayendo consigo la necesidad imperante de una intervención inmediata. Por ello, debe quedar claro con esta investigación que la mitigación de los desafíos éticos de la IA radica principalmente en un compromiso con la transparencia, la responsabilidad proactiva, la minimización de sesgos, y la protección robusta de los datos personales, tanto en el momento en que es diseñada como en su propia implementación en la sociedad, constando siempre con una supervisión humana efectiva.

Referencias bibliográficas

Agencia Española de Protección de Datos. (2020). *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción.*

Beuchot Puente, M. (s. f.). *Los derechos humanos como asunto ético, desde la hermenéutica.* Universidad Nacional Autónoma de México.

Comisión Presidencial Coordinadora de la Política del Ejecutivo en Materia de Derechos Humanos (Copredeh). (2011). *Declaración Universal, Versión comentada.* Gobierno de la República de Guatemala.

Equipo Laboratoria. (2023, 19 de junio). 10 casos donde la Inteligencia Artificial jugó en contra de la diversidad. *Laboratoria: blog de tecnología e inclusión.*

Fernández, E. (1984). Concepto de derechos humanos y problemas actuales. *Derechos y Libertades. Revista del Instituto Bartolomé de Las Casas.*

Grupo de Expertos de Alto Nivel sobre la IA. (2019, 8 de abril). *Directrices* éticas *para una IA fiable. Comisión Europea.*

Guaña-Moya, J. y Chipuxi-Fajardo, L. (2023). Impacto de la inteligencia artificial en la ética y la privacidad de los datos. *RECIAMUC*, *7*(1), 923-930. https://doi.org/10.26820/reciamuc/7.(1).enero.2023.923-930

IEEE. (2023). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

McCarthy, J. (2007). *What is artificial intelligence?* Stanford University.

McCarthy, J., Minsky, M., Rochester, N. y Shannon, C. (1955). *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence.*

Morales Cáceres, A. (2020, 1 de septiembre). *El impacto de la inteligencia artificial en la protección de datos personales*. World Compliance Association.

Naranjo Faccini, R. (2024, 6 de noviembre). *Ejemplos de los dilemas éticos en la protección y análisis de datos*. Skina IT Solutions.

Parlamento Europeo. (2024). *Ley de IA de la UE: primera normativa sobre inteligencia artificial*. https://www.europarl.europa.eu/news/es/headlines/society/20240606STO7904/ley-de-la-ue-primera-noemativa-sobre-inteligencia-artificial.

Russell, S. y Norvig, P. (2021). *Artificial intelligence: A modern approach* (4.ª ed.). Pearson.

Tecnología Educativa. (2024, 11 de octubre). *4.2: Dilemas Éticos en la Protección de Derechos Humanos*. DTI Anáhuac Mayab.

Wolk Soft. (2025, 10 de enero). *La ética en la IA:* ¿Qué nos espera en el horizonte tecnológico de 2025? Amplify Radio.